INTRODUCTORY SYSTEMS AND DESIGN

W. H. Huggins | *Doris R. Entwisle*

THE JOHNS HOPKINS UNIVERSITY

BLAISDELL PUBLISHING COMPANY

A DIVISION OF GINN AND COMPANY

WALTHAM, MASSACHUSETTS • TORONTO • LONDON

Preface

This text has been developed in an attempt to find improved approaches to system theory at the beginning college level. These approaches take advantage of recent advances in precollege mathematics, particularly in algebra and axiomatic modes of thought; the increasing availability of analog and digital computers; and programmed-teaching methods. Noteworthy features are:

- the emphasis of operator graphs, rather than traditional electrical or mechanical circuit elements, in the formulation of system models;
- the illustration of general design principles using "pure" systems composed of only three basic kinds of primitive operators: the scalor, the limitor, and the delayor;
- the incorporation and use of modern algebraic concepts in an axiomatic development of the theory;
- a consistent symbolic notation that distinguishes between a physical entity and the numerical values or functions that describe it;
- a conceptual organization that leads naturally into the study of analog and digital computers;
- many thought-provoking questions and answers that form an integral part of the text and provide self-tutoring instruction.

The insights arising from the development of programming languages for computers have emphasized the importance of algorithmic models in the study of physical systems. Prior to the advent of computers, useful answers could be obtained only by employing the remarkable, but very restrictive, methods of classical analysis to obtain solutions in terms of the classical functions of mathematical physics. These methods usually involve auxiliary assumptions of purely mathematical origin—such as continuity and analyticity—which are necessary for the operation of the analytical machinery but which are, nevertheless, quite irrelevant to the physical problem itself.

Beginning several decades ago with matrix formulations in quantum mechanics, there has been an attempt to reformulate physical problems in discrete, algebraic form so as to emphasize the underlying structure of the problem and suppress these analytical artifacts. The digital computer has accelerated this trend by emphasizing the computational process *itself* rather than the results of the process. Thus, today, one does not ordinarily store tables of functions in the computer. Instead, one stores simple algorithms to compute the values of the functions as they are needed. Because digital computation is inherently discrete, it seems inevitable that the *discrete* calculus* will play an increasing role in the representations of physical systems. It is worthwhile, therefore, to attempt a formulation of system theory with the delay operator as the basic dynamic operator rather than the derivative operator used traditionally. Furthermore, since nonlinear operations are easily incorporated in computational algorithms, an algorithmic approach to system theory can incorporate nonlinear and time-varying elements as well. In short, this book reconstructs introductory material on systems along lines consonant with these modern developments.

In other ways too, the novel features of this book are analogous to some aspects of computer programming: (1) by the use of a carefully designed, mnemonic notational system that is problem oriented, and (2) by the breaking up of the learning materials into small pieces that are carefully sequenced, much as a computer program is sequenced. The notation has many advantages that stem directly from flow-graph methods, but perhaps its chief advantage is a stripping away of all details from the presentation of systems so that the student can readily see the mathematical models in their full generality. The models will not become obsolete. The examples, with their substantive content drawn from several fields of engineering, are overlaid upon the basic substratum of a few simple models. The timelessness of the main content of this book has caused us to engage in extensive research and development prior to its presentation here.

This book evolved in response to a need at The Johns Hopkins University, but this need is clearly not local. Beginning engineering students everywhere require instruction in basic linear models and ways of employing and manipulating them. Here we have used this book with beginning students in all areas of engineering—electrical, mechanical, industrial, chemical, and others. Although symbolism indicating particular variables may change from one department to another, the relationships among the variables—i.e., the mathematical structure—are invariant. Emphasis is therefore upon the *relationships*, rather than upon *things*. This tactic also avoids the difficulty that plagues the teacher of engineering subjects: that of how to fashion physical models that are simple enough to be sound pedagogically but that, at the same time, do not caricature experience by being oversimplifications. The models can be illustrated by either analog or digital computers. These are concrete, and yet they offer a minimum of distracting detail.

This approach takes advantage of recent advances in pre-college mathematics, particularly in algebra and axiomatic modes of thought. The emphasis is on operator graphs, rather than on traditional electrical or mechanical circuit elements, in the

* See J. F. Traub, "Generalized Sequences with Applications to the Discrete Calculus," *Mathematics of Computation 19*, No. 90 (April 1965).

formulation of system models. General design principles are illustrated by using "pure" systems composed of only three basic kinds of primitive operators: scalors, limitors, and delayors. Modern algebraic concepts allow an axiomatic development of the theory. The phrase, "It can be shown," is rarely used in this text. Instead we have attempted to build upon these basic axioms and to develop in a logical and plausible way many of the important concepts needed to understand systems. A natural outgrowth of the emphasis on abstract operators is use of a consistent symbolic notation to distinguish between the abstract physical entity and the concrete numerical values or functions that *we* introduce to describe it.

Introductory courses on engineering systems have traditionally been concerned with particular kinds of things—electrical, chemical, mechanical, and so forth—for the simple reason that until recently these were the only physical entities to use in functioning operational systems. The realization of *pure* systems in the form of analog and digital computers now provides physical realization of universal operators. These may be used to illustrate the engineering design process purely and simply. Previously we had no way of making our symbolic mathematical models come dynamically alive, except by finding various natural processes that would unfold in an isomorphic fashion. Now, the modern computer is a slate upon which we may write our symbols so as to have the consequences of what we write unfold dynamically before our eyes. This is a truly revolutionary development. The invention of the printing press gave man unlimited ability to record information in *static* form; the computer has for the first time added a *dynamic* dimension to this process.

The engineer has lately also turned his attention increasingly toward questions of abstract form, structure, and relationship as he studies problems of information processing and logical system organization. The content of mathematics, just as that of physics, is itself subject to the engineering process. As long as we could exemplify the mathematical content of engineering only by reference to some natural physical system, it was impossible to distinguish between the mathematical and the physical aspects of engineering. For instance, when linear circuit theory was presented as part of physical electromagnetic theory, many students must have been left with the impression that Thevenin's equivalent circuit, delta-to-wye transformations, and the like were consequences of some mysterious properties of nature rather than the direct consequences of linear algebra. Even today, there is ample evidence of confusion between the mathematical model and physical reality. When we talk about *RLC* circuit elements, or springs, weights, and dashpots, and point to symbolic marks on the blackboard, we are indulging in a verbal game. The reference is actually *not* to concrete physical elements, but to an abstracted mathematical system which happens to exhibit idealized relations that correspond approximately to what might be observed in the physical system (should the student ever encounter one!). In the student's mind, deductions from the purely mathematical assumptions are likely to become confused with empirical fact, and *vice versa.*

To avoid this difficulty, we believe that it may be helpful to *factor* the subject matter of engineering into its *physical* and its *mathematical* aspects, distinguishing carefully and deliberately between the two. Thus, system theory may be (and is increasingly) presented as an axiomatic algebra involving idealized elements. These elements are defined to have certain abstract properties and to obey certain relations

when combined into networks or systems. This approach highlights the common mathematical structure of electrical, mechanical, thermal, hydraulic, and other systems without confusing this structure with idiosyncrasies of particular physical devices.

The content of this book has been continually reorganized and presented in lectures over the past several years. Several sets of programmed notes were revised from experience with students and are interwoven into the present volume. This student-directed revision has proceeded along several lines. The major line, as far as this book is concerned, has been through direct experience with students working their way through problems, one by one. At several different times students have been "tutored through" portions of the text. The material was segmented into single-question units, and the authors aided naïve students who were attempting to teach themselves the material. This feedback was supplemented by later trial of the book with large classes of about 100 students each, where students primarily instructed themselves. A few lectures were given on related topics, but a conventional lecture series was abandoned and mainly the material covered was obtained directly from the book by the students. Graduate students served as tutors, each responsible for a small group of undergraduates. The use of the analog computing equipment was developed in these tutorial groups. From their group discussions, the tutors were also able to suggest improvements in the textual material.

A word is in order about the programming of the text. The style is generally linear without being rigidly Skinnerian. Questions are frequently posed, and the student is urged to answer these before going further. The answers appear one or two pages later, just far enough away so the student's eye will not see the question and the answer in the same field. These questions have been carefully prepared and revised on the basis of students' experience, and they are easy to answer if the preceding textual material has been understood. Our research pointed up common kinds of misunderstandings, and explanations of answers now cover these empirically discovered misunderstandings.

Along with these very specific revisions suggested by tutoring of individual students and other microscopic methods, we have conducted a series of experiments that bear on the programming of the text in a more avuncular fashion. We found in one set of experiments, for example, that dual principles in electrical circuit theory tend to interfere seriously with one another when taught in quick succession.* We have tried to implement this finding specifically in the present text. We have also tried to observe the substance of this finding in presenting other bodies of material that are potentially interfering (circuit diagrams and flow-graphs, for instance). Several experiments suggest that specific written responses are differentially helpful to different kinds of students. Based on this finding, our recommendation is that students should try to answer the questions, but if they are not successful, they should consult the answers without further ado.

Our more general research work has prompted us to provide review sections. Students generally favor a programmed approach, if anonymous questionnaires to our

* D. Entwisle and W. H. Huggins, "Interference in the Learning of Circuit Theory," *IEEE Proceedings, 51*, No. 7 (July 1963), pp. 986–990.

own students can be taken as a guide; but an extensive table of contents and review sections also help delineate the broader outlines.

Generally we believe that the most profitable way to use this text is to skim a chapter rapidly, to see where the chapter is pointing, and then to return to the beginning and start through again more slowly, working problems. Some students have found it helpful to prepare flow-graphs of the subject matter for themselves, listing the major concepts in each chapter and showing their interrelationship.

It is a pleasure to acknowledge the help of many colleagues in preparing this book. Chief among these is Dr. Eliezer Naddor, who served for several years on the staff of the course which gave rise to this book. His many suggestions and insights have been illuminating and valuable; in fact, they are so numerous as to preclude specific listing here. While tutoring several experimental student groups through earlier versions of this text, Frederick W. Phelps, Jr. made important contributions. A. Jay Goldstein read the manuscript and offered many helpful suggestions which we have incorporated. The typing of the text and the arrangement of the questions and answers were masterfully handled by Mary Scheller and Betty Howell. The many figures and illustrations are due to the original artwork of James L. Smith. Very special thanks are due to Barbara Bricks, whose diligence and imagination have aided us at every stage of preparation of this book. Overall support of the development of this course has been provided under a grant from the Alfred P. Sloan Foundation. Additional support for the associated research studies was provided under Contract AF 19(628)–263 (Project 7684) with the United States Air Force and by grants from the United States Department of Health, Education and Welfare, Title VII, Projects 10/2 and 1165.

W. H. HUGGINS
DORIS R. ENTWISLE

Contents

I Signals, Operators, and Systems 1

Abstractions—The Way We Think About Things 1
Signals, Observables, and Relations 4
Signals 13
Operators 14
Scalors 16
Comment on notation 17
Dynamic operators 17
Linear operators 18
Delayors 21
Systems of Operators 24
Reversal of algebraic sign of signal 26
More definitions 28
Flow-Graph Representation of an Electrical System 31
Summary 46

2 Signal Flow-Graphs 49

Fundamental Definitions 50
Elementary Transformations 53
Effect of Self-Loops 61
Node Absorption and Graph Reduction 65
Commutative Operators 68

Graph Transmittance (Operator) 74
The Graph Determinant 80
Mason's Loop-Expansion Theorem 89
Mason's General Graph Transmittance Expression 95
Path and Loop Inversion 110
The Converse of a Flow-Graph 119
Appendix A: Proof of Mason's Loop-Expansion Theorem 119
Appendix B: References in Chronological Order 133
Summary—Definitions of Terms 134

3 **Signal Relationships in Physical Systems** 137

A Mechanical System 137
Three-Terminal Electrical Circuits 143
Conductance parameters 144
Resistance parameters 147
Hybrid parameters 147
Relations Between Parameters 152
Cascade connections 155
Feedback connections 167
Operational Amplifiers 170
A vacuum-tube amplifier 172
A transistor amplifier 173
A vacuum-tube amplifier with feedback 176
Cascade connection of amplifiers 183
Operational Amplifier with Feedback 186
Operational circuits made easy 191
Summing circuit 193
Adjustable scalors 195
Summary 204

4 **Operator Graphs** 207

Classification of Operators 208
Systems of Operators 212
Repeated operations 215
Inverse of an operator 216

Approximate realization of inverse of an operator 220
Multiplication and Modulation 221
Servomultiplier 223
Quarter-square multiplier 229
One-quadrant multiplier using LOG and EXP operators 231
Three Basic Operators (Scalors, Limitors, Delayors) 236
Limitors 238
Approximation of static nonlinear operators using limitors and scalors 240
Diode function generators 243
Dead-zone operator 248
Positive and negative limitors 250
Signum operator 255
Hysteretic operator (or flip-flop) 256
Comparison of Nonlinear with Linear Operations 260
Commutative operators 262
Delayors 263
Signal generators 264
Effect of loops 270
Convergence of operator series 282
Stability 284
Summary 297

5 **Weighting Patterns and Filters** 299

The Measurement of Signals 299
A typical measurement problem 300
A filter for obtaining average values 312
Overall instrumentation 319
Design of a compensator 321
The Weighting Pattern of an Operator 327
Weighting Functions 336
Continuous weighting functions 340
Relations to the calculus 347
Continuity 350
Filters in cascade 356
Convolution of weighting functions 359
Weighting Patterns Obtained by Operator Multiplication 369
Summary 375

6 Signal Generators 377

How to Describe Signals 377
Tables of values, graphical plots, and functions 377
Specification of the generating process 381
Weighting patterns again 385
Establishment of initial values 390
The Unit-Impulse as a Primitive Signal 392
Construction of an Operator to Generate Specified Signal Values 398
Component-Signal Generators 409
Power series in D^{-1} 411
Initial values again 414
Partial-Fraction Expansions 419
Partial fractions with repeated factors 431
Partial-fraction expansion for nonreal factors 437
Application of Methods to a Circuit Problem 446
Summary 448

7 Sinusoidal Signals — Their Algebra and Measurement 451

Review of Previous Results 451
Signals of the First and Second Kinds 455
Complex Arithmetic 463
Equality 464
Addition 464
Subtraction 464
Multiplication by a complex scalor 464
Inverse of a complex scalor 465
Complex conjugation 466
Polar representation of a complex quantity 466
Rationalization 469
Real polynomials of a complex variable 475
Measurement of Sinusoidal Signals 479
Steady-state measurements 480
Measurement of signal amplitude 480
Measurement of phase angle 485
Measurement of frequency 489
Complex Transmittances 491

Phasors as Complex Amplitudes 495
Impedance and Admittance as Complex Numbers 496
Complex Power 500
Effective Values of Current and Voltage 505
Summary 505

8 **Frequency-Domain Representations** 509

Why Different Representations? 509
The Null Principle 515
Growing-Exponential Signals 519
Characteristic signals 523
Generation of growing-exponential signals 524
Operator identification by transmittance measurement 524
Truncated decaying-exponentials are not characteristic signals 527
Transmittance Functions 527
Relation to the weighting functions 532
Laplace transforms 533
Uniqueness of transmittance functions 535
Evaluation from experimental data 539
Complex-Exponential Signals 542
Complex frequencies 543
Phasors again 547
Complex transmittances 548
Concluding Remarks 553
Summary 557

9 **Zeros and Poles** 559

The Zeros of a Transmittance Function 560
Iterative Numerical Computations 569
Newton's method 569
Stability and convergence 573
Graphical Interpretation of the Zeros 579
The Poles of a Transmittance Function 583
Natural Frequencies 588
Natural and forced responses 590

Relation to solution of differential equations 590
Solution by operational methods 592
Natural frequencies and resonance 594
Combination of Operators 604
Cascade arrangement 604
Additive arrangement 612
Feedback arrangement 617
Electrical Circuits 619
Thevenin and Norton equivalent sources 624
Ladder networks 626
Signal Decomposition into Exponential Components 631
Generation of periodic signals 635
Expansion of periodic signals into exponential components 636
Spectral representation of d_0 654
Example 660
Summary 663

INDEX 665

INTRODUCTORY SYSTEMS AND DESIGN

W. H. Huggins | *Doris R. Entwisle*

THE JOHNS HOPKINS UNIVERSITY

BLAISDELL PUBLISHING COMPANY

A DIVISION OF GINN AND COMPANY

WALTHAM, MASSACHUSETTS • TORONTO • LONDON

I Signals, Operators, and Systems

Abstractions—The Way We Think About Things

One of the remarkable aspects of the universe of which we are a part is the interrelatedness of each part to every other. This is an aspect of reality that everyone quickly learns to use from the day he is born—it is reflected in our private comprehension of reality, subconscious and conscious, and in the activities and institutions of our collective cultures.

The task of science has been to represent this infinitely complex web of relationships, which we each perceive individually, by a commonly accepted, communicable set of definitions and propositions that can account for as many of these relationships as possible. This task is never complete, nor is it exact, for our conceptual model can represent only certain aspects of the real system that are believed to be important for the purpose at hand. Certain other aspects must be ignored to keep the model from becoming so bulky that it loses its usefulness. For instance, often a straight line will describe the plot of the relationship between two variables over a certain region. In this case it is convenient to speak of "a certain region" rather than to extend the curve so it bends at one or both ends. The equation for a straight line is easy to communicate and to think about—it provides a convenient summary and saves us from having to present a table giving coordinates for a series of points. An analytic model, then, is useful because with only a few numbers we may summarize many numbers.

An essential part of this process is our use of language for naming members of a class. We then regard each member of the class as equivalent in some sense to every other member of the class. To be precise, we should specify other properties or relationships which distinguish each member of the class. For instance, if I were to go to a furniture store to shop for a chair, I would likely be shown a wide variety of contrivances, some made of wood and leather, others of metal and cloth, and perhaps even one or two with an electrical motor to produce restful vibrations. Yet, despite the fact that in size, color, and shape these chairs certainly do not look alike, we still call them all by the same name because we can sit on them. Of course, we may also sometimes sit on the bed, or even on the floor, but then we rarely lie down in a chair (unless it is a Barcalounger) or step

thereon (unless we use it as a stepladder). Classes are useful because we can generalize *within* a class. We assume, hearing that some object is a chair, that it can be sat upon even though we may not have seen the object. Also, classes are useful because different classes can be grouped together into a more comprehensive class—for instance, chairs, tables, and lamps can all be classed as furniture.

This grouping of things, events, and relationships into classes is done so unconsciously and naturally in the early stages of learning a language that we usually fail to realize how arbitrary and sometimes misleading these groupings can be until we compare them with groupings made under the influence of a different culture and language. For instance, in the language of the Hopi Indians, time is not objectified as a linear dimension, so one cannot speak of "an interval of 10 days," and the familiar idea of past–present–future simply does not exist. Also, in that language separate names are given to objects that we could regard as similar. A knife for cutting bread has a different name from that of a knife for cutting meat. What to us seems to be "hard, practical common sense" may make very little sense to a person of another culture—common sense is largely a matter of talking so that one is understood (hence, the adjective "common"). Even within our own society, we may discern C. P. Snow's two cultures—the literary and the scientific—between which common-sense communication has become increasingly difficult.

QUESTION 1.1 Suppose that a human being has grown up completely isolated in the woods (as has happened). This person does not know any language and you would like to teach him. How would you go about teaching him what a book is? (By this, one might mean: the class of objects denoted as books; a book that you have in your hand; or the four-letter word "book." Which meaning would you teach first?)*

The point emphasized here is that everyone is continually classifying things and relationships in subtle and often unconscious ways. These classifications form abstractions from reality on which we later act as if they were the "real" thing. However, what at first may seem most concrete and real may subsequently turn out to be merely that which is most familiar. If every new undertaking were completely unrelated to all that had gone before, engineering—as well as all human activity—would be a hopeless task, for nothing that we had learned previously could be applied in dealing with new problems. Fortunately, this is not the case. In physics, we learn ways of describing and predicting natural phenomena by physical laws that hold *wherever* and *whenever* we apply them. Indeed, the primary goal of science is to discover significant observables that characterize or describe a class of systems and then to express the relationships among these observables in the simplest, most invariant, and most general way. Usually, but not always, each observable may be assigned a numerical value by the use of some specified measurement procedure, thus becoming a physical *quantity*. In this way, a given physical system can be described by the values of the quantities that partially describe it. If the relationships among these quantities are expressed in mathematical terms, we say that the observable aspects of the physical system have been represented by a *mathematical model.* (Other kinds of models, such as physical models, are important too.) If this model correctly predicts the outcome of measurements made under

* When a question ends without special note, there is no answer given in the book.

a large variety of conditions and if it is generally applicable to a wide class of physical systems, it may be elevated to the status of a *physical law*.

Elevation to the status of "law" depends upon our degree of belief in the model or relationship. Complete certainty is impossible because, although 100% of the tests of a model to date have not cast doubt on it, in principle there are an infinite number of tests and we can never complete the testing. Sometimes models yield wrong predictions because the wrong observables are identified with the abstractions—when new identifications are made, the old model may be good. All of this involves abstraction and elimination of that which is unessential or irrelevant.

QUESTION 1.2 What do you think you would do if you had a model that worked well in many cases, but then you found a case that didn't "fit"? Have you been able to define "book" for the person described in Question 1.1 so that he can distinguish between books and magazines? Between books and sheet music?

One of the noteworthy consequences of describing a physical system by a mathematical model is that the mathematical formulation forces us to make abstractions and hence leads toward a common language for comparing various systems and displaying common features and significant differences. Once the physical observables have been translated into proper numerical form by appropriate measurement procedures, systems which appear on the surface to be quite unrelated physically may nevertheless have identical mathematical structure. Such systems are said to be *analogs*. We may describe this situation by saying that analogous systems are *physically* different but *mathematically* similar. In fact, all linear stationary systems (which form the main subject of this book) are mathematically equivalent to each other in the sense that the mathematical theory of one will serve to illustrate the theory of all. This is why we are discussing "system theory" rather than circuit theory, mechanics, or some other topic.

It should be noted, however, that the *same* physical system may be described by *different* mathematical models. This is so because any model represents only certain selected aspects of the physical system; by selecting different aspects, one obtains different models. Furthermore, even the *same* set of aspects of a given system may be represented by a variety of different mathematical models. Thus, whether two systems are analogous depends not only on the systems *per se*, but also on how we choose to think about them—and this involves *nonmathematical* considerations, such as physical significance, value, and meaning. In subsequent chapters, the emphasis will be on the mathematical model rather than on the many detailed physical considerations which must precede its establishment. These physical considerations are of the utmost importance, for they provide the link between mathematical theory and reality. Although it has been customary to teach the physical aspects of a system simultaneously with its mathematical theory, in this book these physical considerations have been deemphasized, *not* because they are unimportant, but rather because we believe that the traditional presentation has resulted in confusion of the physics with the mathematics and has obscured the underlying mathematical similarities among different kinds of physical systems.

To avoid this difficulty, we shall illustrate the mathematical aspects using a computer. A computer—digital or analog—may be regarded as a "pure" system in which many of the physical considerations have been deliberately suppressed so that one can focus one's

attention on the purely mathematical aspects of the system. Until recently, pure systems existed only within the human mind as constructs designated by mathematical symbols on paper. However, the modern electronic computer now provides for the first time a physical realization of a pure system which can exhibit in full dynamic detail the logical consequences of its design. In this book we shall develop many of the ideas widely used today in engineering systems, using a pure system that exhibits the essential mathematical properties but which is free of confusing physical detail that is often irrelevant to the mathematics under discussion. It allows us to get a "perfect" physical model to represent a mathematical model. Then it is easy to explore the consequence and workings of the mathematical model using the computer rather than pencil and paper.

Signals, Observables, and Relations

In the previous section, we observed that the mathematical model of a system represents the relations among the numerical quantities that describe the system. What these quantities should be, and how to measure them, are difficult questions that fall outside the domain of ordinary mathematics. In fact, it is here that we often find a host of difficulties that can be resolved only by reference to the physical system itself. Even this may not be adequate for, as Norbert Wiener has observed,* "Things do not in general run around with their measure stamped on them like the capacity of a freight car; it requires a certain amount of investigation to discover what their measures are.... What most experimenters take for granted before they begin their experiments is infinitely more interesting than any results to which their experiments lead." Because these questions are largely evaded in our treatment of pure systems, it is reasonable that we discuss them briefly in the remaining sections of this chapter so as to preserve some perspective on the overall problem.

The observables with which we shall be primarily concerned are those which describe the state of the system at any instant of time. The value of each observable will be a function of time. We shall call such observables *signals*.

Originally, the word *signal* meant some form of communicative sign, but because of its widespread use in radio communication to denote impulses, radio waves, and telemetered messages of all kinds, the meaning of the word *signal* has broadened until it now includes practically any time-dependent observable associated with some system. Examples of signals that may be observed in an electrical system are the voltages, currents, displacement of electric charge, magnetic flux, and associated mechanical forces. In mechanical systems, the velocity, displacement, and acceleration of various parts of the system, as well as the forces and torques associated therewith, may be regarded as signals.

From the foregoing remarks, one might be tempted to conclude that almost any measurable quantity associated with a system could be called a signal. However, some measurement processes yield values that express the coefficient of proportionality between two signals. This kind of measurement is often needed to describe quantitatively some *property* or quality of the system. Such *derived* quantities which are determined

* Norbert Wiener, who was the originator of the science of cybernetics, was a mathematician on the faculty of the Massachusetts Institute of Technology. The quotation is from "A New Theory of Measurement: A Study in the Logic of Mathematics," *Proc. London Math. Soc.*, Ser. 2, *19* (1920), pp. 181–205.

from more than a single measurement are usually called *parameters.* Examples of parameters in an electrical system are inductance, resistance, and capacitance, which describe the relation among the voltage and current signals associated with the three different kinds of circuit elements. This distinction between signals and parameters seems quite reasonable in constant-parameter systems in which the values of the parameters do not change with time. But in variable-parameter systems, the distinction becomes increasingly arbitrary and, in fact, is largely dependent on one's point of view.

The kinds of difficulties that are encountered in selecting a set of observables may be illustrated by the following story.

John Jones is beginning the last lap of his journey west by automobile after staying overnight in a small town. Just as he reaches the edge of town he runs out of gas, but fortunately he is able to reach a nearby service station just as the motor coughs and stops. He asks the attendant to fill his tank. While this is being done, he amuses himself by thinking of different ways in which he might determine the amount of gasoline that has been added to his tank. Several possibilities occur to him:

1. The user's manual which came with his car stated that his tank has a maximum capacity of 20 gallons. Since the tank was initially empty, the attendant will have pumped in 20 gallons.
2. A simple calculation involving the amount of gasoline used and the total distance traveled (as read by the speedometer) shows that his car had averaged 22 miles per gallon of gas. Hence, he could continue on his way until he again runs out of gas. Then, dividing how far he had traveled (as indicated by the difference in the mileage readings of his speedometer) by 22 miles per gallon should yield the number of gallons that had been added at the last filling.
3. He could (if he had only thought of it in time) weigh the car just before and just after getting the gasoline. From the difference in the before-and-after weights of his car, he could determine the amount of gasoline in gallons (if he could only remember the weight of one gallon of gasoline).
4. He could divide the amount of money that he paid the attendant by the price of one gallon of gasoline.
5. He could measure the shape of the tank and the depth of the gasoline in the tank by inserting a stick. Then, by using some geometry he might calculate (if only he could remember the formulas) the volume of the gasoline that would fill the tank to the measured depth.
6. He could, after turning on his ignition switch, read the gasoline indicator on the dashboard of his car. Since it has a scale that is labeled in fractions of a full tank, multiplication of this scale reading by 20 should yield the number of gallons in the tank.

QUESTION 1.3 Which of these possibilities would be most likely to yield the most accurate estimate of the amount of gas that was received? Discuss the reasons why you prefer this possibility over the others.

QUESTION 1.4 Which of these methods of determining the amount of gas in the tank would you consider to be a measurement?

QUESTION 1.5 Each of these ways in turn makes use of other measurements. Which of these might you consider to be the value of a signal? A parameter? A property?

QUESTION 1.6 Compare possibility 2 with possibility 6 and discuss in what ways they are similar and in what ways they may differ. In particular, discuss what you think is meant by "the calibration of a measurement procedure."

QUESTION 1.7 Suppose that, instead of determining the increase in weight of the car as in possibility 3, it were possible to weigh the storage tank from which the pump obtains its gasoline. Could measurement of this weight before and after delivery of the gas provide a seventh possibility?

QUESTION 1.8 Suppose that the meter in the very old-fashioned gasoline pump consists of a little propeller in a cylindrical pipe through which the gasoline flows, thus causing the propeller to rotate at an angular velocity that is directly proportional to the velocity with which the gasoline moves through the pipe. The propeller is connected through a gear train to an indicating pointer (rather like a clock mechanism) evident on the outside of the pump. The scale associated with the pointer is marked off to read the number of gallons delivered.

Next, consider the second possibility described above, but suppose that instead of determining how far the car has gone before it again runs out of gas by directly reading the mileage indicator, John Jones now reads his speedometer at intervals of one minute during the last part of his journey, and he records these successive readings. How might he use these data to determine the amount of gas? Is there any resemblance between the process used by Jones and the measurement procedure used by the designer of the gas pump?

Perhaps in thinking about these questions you have already discovered that the ideas illustrated by this rather silly story are not at all trivial or insignificant. In fact, most of the important considerations that enter into our choice of observables are involved here, and we shall discuss some of them briefly.

Relations • An observable has meaning and usefulness only because it is *related* to other observables. Hence the choice of the observable must depend primarily on the particular *relations* that we wish to establish.

Two observables may be related *directly* or *indirectly*, depending on whether the relation involves *intermediate* observables. For instance, it was observed that John Jones subsequently had for supper only one hamburger and two cups of coffee. He would have eaten more except that he would have had to part with the silver dollar that he received in change when he purchased the gasoline. Now most people would regard the relation between what Jones ate for supper and the amount of gasoline in his tank as indirect, and for good reason. The relation might have been altogether different if the price of gasoline had been only slightly different; if he had been less tired and of better appetite; if he did not have a young son; if he had a hole in his pocket; etc. The relation between the place along the highway where Jones would run out of gas (possibility 2) and the amount of gas placed in the tank is *indirect* because the process whereby the

automobile engine converts the fuel into a force which in turn propels the car along the road must certainly intervene. Yet we might wonder whether, in principle, this relation is any more indirect than the relation between the pointer of the gas gage and the amount of gas received (possibility 6). In the first instance, the car itself is the pointer and the road the scale; in the second instance a mechanical lever is actuated by a float on the surface of the gasoline which in turn changes the resistance of an electrical circuit, thus allowing more current to flow through a heating element that produces a motion of the gasoline-gage pointer as a consequence of the change in temperature of a thermally sensitive spring. Thus, both relations can be described using intermediate relations.

Yet, we are inclined to think of the relation between the reading of the gas gage and the number of gallons in the tank as being much more direct than any of the other relations mentioned above. We may discern several reasons why this should be so. First, there is the feature of *temporal simultaneity*. The gas gage produces an indication of the amount of gas in the tank almost as soon as it was put there, whereas most of the other possibilities involved relations between events happening at different times. (This remark would probably be incomprehensible to the Hopi Indian.)

Second, the scale of the gas gage is labeled so as to refer directly to the amount of gas in the tank. This *implies* a *direct* relation and we are accordingly likely to think of it as such even though on more careful examination we find that it is described by many intervening observables. Furthermore, it should be noted that the speedometer could be calibrated in *gallons* of gasoline used rather than in *miles* traveled, and that this might provide a *more accurate* indication than the typical gas-tank *indicator*. It appears that we are here concerned more with how we think about the relations than with fundamental differences.

What then is the most important characteristic that distinguishes a direct from an indirect relationship? We suggest that it is primarily a matter of its *predictability*. If the relationship may be expected to exist quite independently of the relationships with other observables in the system, we may say that the two observables are directly related. This is clearly the case with the gas gage, for despite the mechanical, electrical, and thermal observables associated with its operation, there is a more or less predictable relation between the pointer reading and the amount of gas in the tank. This is clearly *not* the case in the relation to Jones's supper, for that relation could not have been predicted (although it perhaps could be *explained* provided we obtained additional information about the system). Thus, a relation may be classified as either *deterministic* or *probabilistic* (i.e., random) depending on how *certain* we are of its occurrence. This classification is largely a matter of degree, and relations may fall anywhere between the completely deterministic and the completely random.

QUESTION 1.9 Is the amount Jones paid for the gasoline and the amount of gas actually placed in his tank described by a deterministic or a probabilistic relation?

It is commonplace to think of relations in terms of *cause* and *effect*. For instance, almost everyone would agree that the gas gage points to "$\frac{1}{2}$" "because" the tank is half full of gasoline. At least, when additional gas is added to the tank the gage reading is always observed to increase. In contrast, if someone were to increase the pointer reading of the gage by moving a small magnet across the cover glass, it would be most surprising if the

gasoline in the tank were increased thereby! The relation between the amount of gas in the tank and the gage reading clearly seems *asymmetrical*—the gas in the tank is the "cause," and the meter reading is the "effect." It certainly appears to make little sense to reverse the relation and assert that the gage reading is the "cause," and the gasoline in the tank is the "effect."

To return to Jones's odyssey, after leaving the gas station, Jones decides to trust his gas gage. He drives all day. As he reaches the town where he plans to spend the night, he observes that the gas gage indicates that his tank is empty, and in response to this indication he finds a service station and has the tank refilled.

QUESTION 1.10 For this situation, discuss which is the "cause" and which is the "effect." If you believe that the reading of the gage "caused" the replenishment of the gasoline in the tank as the effect, are you now admitting that the earlier cause–effect relation (on which I thought we had agreed) is no longer valid? If it is not valid, why not? (But do not tell Mr. Jones that it is not valid, or he will have his gage replaced by a new one for which it is!)

You perhaps have found that the notions of "cause" and "effect" are not nearly so clear-cut as they at first seemed. The reason for this difficulty can be illustrated by a simple diagram (see Figure 1.1). The upper arrow, drawn from the box labeled "Gas in

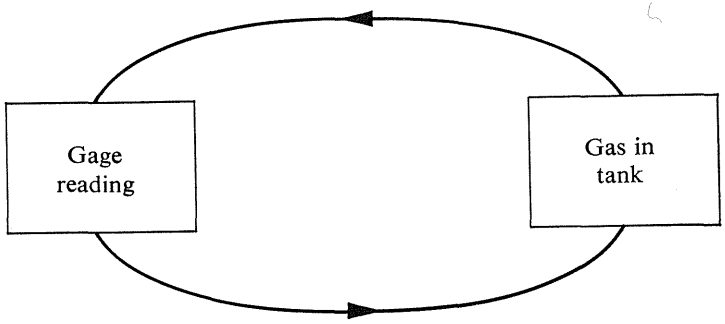


FIGURE 1.1

Tank" to the box labeled "Gage Reading," denotes the mechanical-electrical-thermal mechanisms whereby changing the surface level of the gasoline causes a corresponding change in the gage reading. The bottom arrow, drawn in the opposite direction, denotes the other mechanisms (including Jones and the service-station attendant) whereby the gage reading below some point "caused" a corresponding change in the amount of gas in the tank. As this diagram illustrates, we can speak of cause and effect only when we suitably *isolate* and *restrict* parts of the system. In this example, complete isolation is possible in that if we chose to do so, we could arrange the system so that the gage reading in no way affects the gas in the tank. Likewise, by turning off the ignition, we could ensure that the amount of gas has no effect on the gage reading. But note that the separation of these relations in a meaningful way involves making special physical arrangements so as to *force* the modified system into agreement with our patterns of thought. When these

relations form one or more *closed loops*, as in the diagram above, the relations between the observables are altered in remarkable ways. The process of "separating" a system into its "component parts" between which various *causal* relations may be said to exist is an important part of constructing a *feedback system* model, about which we shall have much more to say in later chapters.

Before we leave this discussion of cause and effect, we should ask whether there is any aspect of our environment that does definitely demand an asymmetrical relation. Many relations are asymmetrical *only* because we choose to think of them as such. Sometimes this choice is so strongly conditioned by traditional usage that we may find it virtually impossible to consider that the inverse relation may be equally valid physically. For instance, Norman Robert Campbell, who was one of the major contributors to the philosophy of science, was able to think of Ohm's law as a symmetrical relationship in which either the voltage or the current could be considered as the cause, and the other as the effect. But for Hooke's law, he could only accept the force "acting" on an elastic member as the cause of the "resulting" elongation. That the elongation might, instead, be regarded as the cause resulting in a force, he held to be "obviously" fallacious (see p. 61 of his *Foundations of Science—The Philosophy of Theory and Experiment*, Dover Publications, S372). Similarly, most people today find it much easier to think of sources of voltage, such as a battery, than of sources of current, although in my opinion both are physically acceptable ideas. Thus are we bound by familiar ways of viewing things around us to regard the less familiar as "unnatural."

Despite the difficulties of clearly defining what we mean by a "causal" relation, we do find a particular relation among observables which automatically implies *noncausality*. This is the *temporal-ordering relation* expressed by "before" or "after." Consider these events, A, B, and C, which "happen at different times." As you know, the ordering relation is *transitive*, which assures us that if A occurs before B, and B occurs before C, then A surely occurs before C.

QUESTION 1.11 If A occurs *before* B and C occurs *after* B, does C *surely* occur after A?

An essential requirement of *the notion of causality is that the "cause" must precede the "effect."* This requirement is consistent with our intuitions, and, for many purposes, may be used as a test for *non*causality. For instance, if A occurs *after* B, then A *cannot* be the *cause* of B.

QUESTION 1.12 Reexamine Question 1.10 in the light of the foregoing discussion. Can the temporal-ordering test be used to clarify the noncausal relations in this question? If a relation cannot be shown to be *noncausal*, is it then necessarily *causal*?

Unfortunately, although we may test for *non*causality, difficulties remain when we attempt to demonstrate causality. Suppose that event A causes both an event B and an event C, and that event C is observed to occur after event B. Furthermore, assume that the occurrence of A was *not* observed. Our test then correctly assures us that C is *not*

the cause of B, but it leaves in doubt whether or not B may be regarded as the *cause* of C. (In the beginning, there was the A which "caused" all that followed. . . .) We shall not attempt to pursue the notion of causality further, but shall summarize this discussion by emphasizing that the notion of causality is intimately related to "the passage of time" and how we choose to *think* about events. As a matter of fact and strictly speaking, it is impossible to prove cause and effect. It is nevertheless helpful to think of events as causally related.

Let us return to the question of what considerations govern the choice of the observables that we select to describe a system. We have already noted that this choice depends on the particular relations with other observables that *we* may wish to establish, and that this choice therefore requires an understanding of the physical system itself.

Other factors that govern the choice of an observable are its *invariability*, its *generality*, and its *additivity*, each of which we shall now discuss briefly.

By *invariability*, we mean that the *significance* of the observable is the same at any location in the universe and at any time. (Of course, the *value* of an observable may depend on location and time, even though its *significance* does not.) Perhaps one of the most important observables is physical *mass*. Jones made use of the invariability of mass when in possibility 3 he assumed that the weight of the gasoline removed from the storage tank in the filling station would be identical to the weight of the gasoline added to his car. Other observables of fundamental importance in physics are *length*, *time*, *electric charge*, and the related observables such as *area*, *volume*, *velocity*, *acceleration*. Each has a physical significance that is independent of *where* and *when* the observation takes place.

By *generality*, we mean that the observable has physical significance in connection with a wide variety of physical systems. Clearly, most of the observables mentioned in the preceding paragraph are not only invariable, they are also very general, and they arise in the description of physical systems of all kinds.

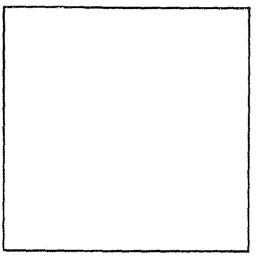
Finally, by *additivity* we mean that the observable is capable of being related to the real numbers in such a way that numerical addition of these values of the observable corresponds to some meaningful, physical process of composition. For instance, if Jones had initially asked for only 5 gallons of gasoline, but then after that amount had been placed in his tank, he changed his mind and asked for an "additional" 12 gallons, the ultimate *physical state* of his gas tank would have been essentially the same as if he had asked for 17 gallons at the outset. Furthermore, the same final physical state would have been achieved had he first asked for 12 gallons and then 5 gallons. In other words, an *additive* observable exhibits in some physically meaningful sense the *associative* and *commutative* properties of ordinary numerical addition.

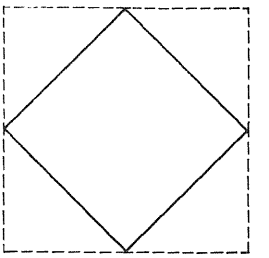
QUESTION 1.13 If Jones had been traveling in Europe instead of the United States, he would probably have asked for some number of *liters* of gasoline. Do you believe that the possibility of describing the quantity of gasoline in terms of different *units*, such as gallons and liters, is necessarily associated with the property of additivity? Why?

QUESTION 1.14 If you were to add two quarts of water to two quarts of alcohol, would you end up with exactly one gallon of the mixture?

QUESTION 1.15 If you were to add 500 grams of water to 500 grams of alcohol, would you end up with exactly 1000 grams of the mixture?

QUESTION 1.16 *Length* is not an additive observable for the amount of paper contained within these squares, but *area* is. What process of *physical* composition can you find to show that the square on the left has twice the area of that on the right?





QUESTION 1.17 If you were to cut out from the paper each of the above squares, you could weigh each of them to determine its mass. Would the *area* and the *mass* of each square be related in any way that is at all *analogous* to Ohm's law, Hooke's law, or both?

Throughout this entire chapter, we have emphasized the fact that it is the *relations among observables* that are of ultimate significance. For instance, the relation between the sizes of the squares considered in Question 1.16 is as the relation of 2 to 1, regardless of *units* used for expressing the *area* of the two squares. When an observable can be expressed quantitatively so that its relative value is proportional to the relative magnitude of some physical attribute, it is said to be *measured on a ratio scale.* Practically all the observables with which we are concerned in this course will be of this type. In fact, we may assert that as a general principle, the intrinsic, absolute meaning and significance of physical observables is expressed *only* by their *relative* magnitudes. Relative magnitudes are independent of the arbitrary units (such as cents, dollars, gallons, ounces, etc.) used as their measure.

The several examples given thus far have also illustrated that when two observables connected with the same system are observed to be directly proportional, we may use this *coefficient of proportionality* to describe some physically meaningful property of the system. For instance, if in Question 1.17 you formed the ratio of the *weight* of the paper squares to their *area*, you would find that this ratio would be substantially the same, whatever the size of the square. It therefore expresses a *property* of the piece of paper from which the square was cut. Since the same ratio is obtained (ideally), whatever the size of the square, this property exhibits *invariability*.

Furthermore, by assembling squares which are sufficiently tiny, we can construct *any* figure, such as a circle. Hence, this property is also *general* and depends neither on the *size* nor the *shape* of the figure drawn on the paper. This property is called a *density*. It expresses the ratio of mass to area for any figure. The total mass of any figure cut from this paper may then be found by multiplying its area by the density of the paper. In our study of systems, we shall encounter many such coefficients of proportionality between

different pairs of observables. It is because of this possibility of finding such coefficients of proportionality that we are able to construct a useful mathematical model of many of the physical systems that will concern us in this book.

Mathematical models may involve nonlinear relations; indeed, many physical systems cannot be adequately described by simple linear relationships. However, the analysis of nonlinear models is usually very difficult. The theory of linear models is simple (and hence useful) because *a linear function of a linear function is itself a linear function.* The kinds of relationships that can arise are therefore much more limited than when nonlinear relations are present.

Finally, it should be noted that not only are we sometimes interested in the *ratios* of certain pairs of observables, but we also may need to form their *product.* Products of the values of certain related pairs of observables play a vital role in the study of engineering systems. It is found in physics that to measure the *work* done on a system requires the simultaneous measurement of *two* observables. For instance, work is defined as "a force acting through a distance," and the value of the work is expressed by the value of the

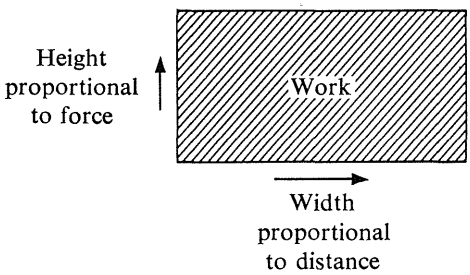


FIGURE 1.2

force multiplied by the value of the distance (under the assumption that the force is constant and acting in the direction of the displacement). If we think of the distance traversed as analogous to the width of a rectangular figure and the constant force as analogous to the constant height of the rectangle, then the work done is analogous to the *area* of the rectangle. The same amount of work could be done by a small force acting through a large distance, or a large force acting through a small distance.

The significance of work is that it is a form of *energy*, which, like mass, is conserved; and that the primary observables in physical systems are those which when taken in suitable *pairs* (like force and distance, or voltage and electric charge) provide a measure of the *energy* of the system. Although energy is additive, it should be noted that it is related to such observables as force, voltage, and the like in the same way as *area* is related to *length.* Because work (and energy) involves *products* of pairs of signals, it demands different treatment from the signals themselves. Hence, we shall not ordinarily think of energy as a signal, although it is closely associated with signals.

Let us summarize some of the main points of this section. We have discussed several factors that govern the choice of the observables used to describe a system. These depend, first of all, on what we are interested in and how it is related to the many observables of the system. We have discussed several attributes of the relations between observables and have concluded that the most important observables exhibit relatively *deterministic,*

invariable, and *general* relations to one another. Furthermore, we shall be particularly concerned with *additive* observables, which can be so related to real numbers that numerical addition of the values of the observable corresponds to simple, physical operations. It is the important consequences of additivity that permit us to manipulate and "solve" the quantitative mathematical models of the physical system.

Signals

The value of an observable depends on *where*, *when*, and *how* it is measured. For instance, if you were to measure the temperature of the air in the room where you are sitting, you would find that the temperature near the ceiling is probably higher than that near the floor. Even when measured at the same location, the temperature would probably be lower in the morning than in the afternoon. And if you were to measure the temperature simultaneously with two different kinds of thermometers at the same location and time, it is possible that you would obtain different temperature values, for one thermometer may be more sensitive to the radiant heat of the sunlight entering the window. Also, a mercury thermometer might be very slow and sluggish and yield a temperature reading which is a kind of average temperature over the previous minute or so, whereas a "hot-wire" or thermistor thermometer would yield an "instantaneous" temperature whose value might fluctuate considerably from instant to instant as thermal convection currents bring air of different temperatures past the thermometer.

Let us examine the role of the thermometer. It produces a numerical scale reading at any instant t which describes some aspect of the room temperature. This aspect depends, as we have seen, on where the thermometer is placed in the room and also upon its own special characteristics. But in any event, it enables us to associate a *numerical value* with the temperature, T, of the room. This value depends on the time, t, and it is convenient, therefore, to denote the *value* of the scale reading at the instant t by $T(t)$. Of course, the numerical value also depends on the *kind* of thermometer and *where* it is placed. For each different location and for each kind of thermometer, we will, in principle, obtain a *different function* $T(t)$ relating the time t to the scale reading. Consequently, we must distinguish between different *locations* and/or different *thermometers* by a subscripted index that uniquely designates each of the various arrangements being considered. For instance, the temperature as measured 1 ft from the ceiling near the southeast corner of the room might be designated by T_1; the temperature 1 ft from the floor in this same corner might be denoted by T_2; the temperature in the middle of the room by T_3, etc. The value of the temperature T_1 at a particular instant t would be denoted by $T_1(t)$; the values of the temperature T_2 at this same instant (but at a different location) would be denoted by $T_2(t)$; etc.

The symbols T_1, T_2, etc., denote different *signals*. Each symbol represents some *observable aspect* of the room temperature to which a numerical value may be assigned at each instant t by a specified process. A signal, such as T_1, has a *value* which depends on the time t at which the value is read. Hence, the numerical value of T_1 at the instant t may be written as $T_1(t)$. However, T_1 by itself is *not a number*; rather, it denotes the entire physical process whereby some *aspect* of the *physical temperature* at the upper southeast corner of the room is mapped onto the *single*, *numerical scale* of a particular thermometer. The numerical reading on that scale at the instant t is then $T_1(t)$. Thus,

a signal is a physical observable to which a unique, single number may be assigned at each instant of time. The meaning and physical significance of the signal is defined by the measuring process.

QUESTION 1.18 Suppose that we wish to determine "the temperature" of the room by averaging the temperature near the ceiling and the temperature near the floor of the room. We could install two thermistors which yield two electrical signals T_1 and T_2 and then combine these two electrical signals in a suitable circuit to get a *single* signal T, where $T = \frac{1}{2}(T_1 + T_2)$. What does this equation imply about the relation between the values of the three signals T, T_1, and T_2? (Answer)*

The difference in the meanings of Equation 1.1 and Equation 1.2, given in the answer to Question 1.18, should be emphasized. The first equation, 1.1 refers to a physical composition of two signals to obtain a third physical signal. Equation 1.2, on the other hand, refers to a relation among the three *numerical* values that describe these three physical signals at each instant t, and pertains to the mathematical model.

It is unfortunately a common practice to use the same equation for stating a relationship about *physical* signals as well as about their *numerical values.* The fact that these relationships are of similar form in no way minimizes the fact that relations among numbers are fundamentally different from relations among signals. A full appreciation of this difference is vitally important, particularly for the engineer, who must deal with the real world rather than a mathematical model thereof. Because of this ambiguous use of our mathematical notation, it is very easy to confound numbers with that which the numbers represent. *Remember, you can take the logarithm of the temperature, but not the temperature of the logarithm!* (For instance, the voltage delivered by two batteries connected in series is additively composed of the voltage of each battery. Here *physical voltages*, not *numbers*, are being "added.")

Operators

In physical systems, it is useful to think of one signal as the "cause" of another signal. For instance, the elongation x of an ideal spring is related by a simple proportionality to the force f applied thereto:

$$cf = x. \tag{1.3}$$

In general, the applied force f may change with time. Equation 1.3 then states that at any instant t the value $f(t)$ of the force is related to the value $x(t)$ of the displacement by the numerical equation

$$cf(t) = x(t). \tag{1.4}$$

The relation expressed by Equation 1.4 exhibits two special, important properties. First, it is a *static relation.* Even though the elongation of the spring may be different at different times, this equation asserts that the value of the elongation at any instant depends *only* on the value of the force at that *same* instant. Second, it is a *linear relation.*

* When a question ends with the note (Answer), look for the answer on the left-hand page *following* the question or shortly thereafter.

If we were to make a plot on cartesian coordinates of the points $(f(t), x(t))$, for several different values of t, these points would lie on a straight line that passes through the point (0, 0) with a slope c.

The assumption that a static relation exists between two signals is often not justified in physical systems. For instance, suppose that a small weight is fastened to the end of the spring previously considered, and the force f applied in the downward direction to the weight. What now would be the relation between f and the elongation x? Certainly, it requires little imagination to see that the displacement and force are no longer proportional in the sense of Equation 1.4. The value of the displacement at any instant is now determined by the value of the *force at earlier instants of time as well.* For example, a force applied only during some 1-sec interval will cause a displacement signal which has nonzero values a long time after the value of the force has been reduced to zero. However, it cannot cause a displacement *prior* to the time at which the force was applied.

The relationship among signals in a physical system is thus seen to be of a more general kind than a simple relationship among the signal values at the *same* instant. Instead, the value of a signal at a given instant will generally depend on the values of other signals at *all prior* instants. Thus, the weight is displaced up or down "because" at some *earlier time* a force acted on the weight–spring system. Systems for which the values of the observables at any instant depend on the *prior history* of the observables are called *dynamic systems.*

QUESTION 1.19 Is a physical spring truly a static system? Can you describe any other physical systems which are truly static? In what sense may some dynamic systems be "approximately" static?

In a dynamic system, a simple proportionality between two signal values at each instant is not adequate for describing the signal relationships because the value of a signal at any instant may depend on the values of the other signals at earlier instants. Even so, it is useful to preserve the notion of cause and effect and to have some way of showing that the force f is the "cause" of the displacement x. In the sequel, we shall portray the cause–effect relation by a signal flow-graph, as in Figure 1.3.

FIGURE 1.3 *A simple signal flow-graph.*

In the flow-graph of Figure 1.3, each signal is represented by a *node* (a small circle or dot). The dependency of one signal on the other is represented by a directed *branch* (a line) drawn from one node to the other. The branch carries an arrow pointing away from the "cause" and toward the "effect."

In using flow-graphs, we will often find it helpful to think of signals as "flowing" through branches. However, as the signal passes through each branch it is operated on by the branch and transformed *into a different signal.* The branch is therefore labeled by a symbol, such as H in Figure 1.3, to denote the particular operation that is *performed* on the "input" signal f to yield the "output" signal x. This is, of course, merely a notational convention for picturing relationships among signals. These pictures will

enable us to see the relationships more clearly, particularly when several interacting signals are involved.

The relation shown in Figure 1.3 may also be expressed algebraically using *operator* notation. We may denote the signal that results from H operating on f by a *new composite* symbol, fH. For the time being, this new symbol, fH, should be regarded as a

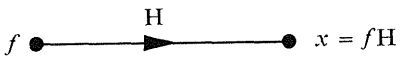


FIGURE 1.4

synonym for x. (The fact that the name fH for the displacement signal is composed of the letters "f" and "H" is merely a useful convention for recording the genesis of the signal fH. It in *no* way implies *multiplication*, just as "CAT" does not mean "C" times "A" times "T.") The value of the signal f at time t will be denoted by $f(t)$, and the value of the signal fH will be denoted by fH(t), just as before. Thus, using operator notation we may express the relationship between the force and the displacement as,

$$f\mathrm{H} = x, \tag{1.5}$$

where, again we emphasize that fH and x are two signals to each of which a value may be attached at any instant t:

$$f\mathrm{H}(t) = x(t). \tag{1.6}$$

Scalors • It is important to compare Equation 1.6 with Equation 1.4, which applied for the ideal, "weightless" spring, $cf(t) = x(t)$. The "ideal" spring is typical of a class of physical operators which in operating on any signal merely modifies it by a constant factor of proportionality. Let A denote such an operator. Then, for any signal f,

$$f\mathrm{A}(t) = af(t), \tag{1.7}$$

where a is a numerical scale factor (i.e., a *scalar*), which relates the value of the modified signal to the value of the original signal. Any signal f, for which Equation 1.7 applies, is called a *characteristic signal* (or an *eigensignal*) of the operator A, and the constant of proportionality a is called the *characteristic value* (or *eigenvalue*) associated with the signal f. A physical operator for which *all* signals are characteristic (so that the effect of the

ANSWER TO QUESTION 1.18 You will recall that our observables were to be additive so that numerical addition of their values corresponds to a physical composition in some sense. Thus, by the statement

$$T = \tfrac{1}{2}[T_1 + T_2] \tag{1.1}$$

we mean that the signals T_1 and T_2 have been combined to obtain a new signal T, such that

$$T(t) = \tfrac{1}{2}[T_1(t) + T_2(t)]. \tag{1.2}$$

At each instant t (within the time interval over which the composition is to occur), the value of T is equal to $\frac{1}{2}$ the sum of the values of T_1 and T_2 at that same instant.

operation upon any signal is simply to yield a new signal which is everywhere proportional by the *same* factor a to the old), we shall call a *scalor*.

A scalor is the simplest type of physical operator. It transforms any signal into another signal whose value at any instant is simply some constant scalar factor times the corresponding value of the first. It is customary to place this scalar multiplier on the *left* of the signal. Thus, the signal $2f$ has a value at any instant that is twice the value of f at that same instant. Evidently, a scalor is a *static* operator. It is *memoryless*, since the value of the "output" signal at a given instant does not depend on the values of the "input" signal at earlier instants. Furthermore, it is clearly *linear* since a plot of the output-versus-input value at any instant will yield points that lie on a straight line of slope a and Equations 1.9a, b, c are satisfied. Evidently, the scalar coefficient a completely

FIGURE 1.5 *A scalor is described by its transmittance.*

describes the scalor. It is called the *transmittance* of the operator, because it describes the proportionality factor by which any signal is transformed as it is transmitted through the scalor.

Comment on notation • In most texts, the same symbol is used to denote both the *physical* scalor and the scalar value of its transmittance. In this book, however, we shall distinguish between physical operators and the numerical coefficients that describe their effects by placing the operator on the *right* of the signal on which it operates; whereas the numerical coefficient will be written on the *left* of the signal.* To further emphasize the distinction between physical operators and scalar coefficients, we may denote operators by Roman letters (usually capitals) and scalars by italicized letters (usually lower case). This convention is illustrated by the equation, $f\mathrm{A} = af$, which would be satisfied by *any* signal f when A is a scalor having a scalar transmittance a.

Dynamic operators • A simple illustration of a dynamic operator is provided by the speedometer in Jones's car. It exhibits two signals: the mileage m, and the speed s. At any instant t, the value of the mileage is $m(t)$, and the speed is $s(t)$. Both signals arise from a common physical process—the motion of Jones's car—and they are clearly related to each other. If we denote the operator which relates s to m by the name RATE, we may represent this relationship in both flow-graph and algebraic forms as shown in Figure 1.6. By our notational convention, the indicated value of the speed at the

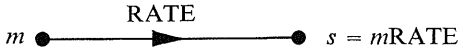


FIGURE 1.6 *Mileage rate gives speed.*

* Unfortunately, it is common practice in most texts today to place both operators and scalars on the left of the signal. This results in confounding the different meanings of the two. It further leads to an awkward reversal in the natural order of writing sequences of symbols when several operators are applied in succession. By appending the operator on the right rather than on the left of the signal, we adopt an algebraic notation that distinguishes *numerical scalars* from *physical operators* and is also in better accord with the subscript conventions used in signal flow-graphs.

instant t may be denoted by either $s(t)$ or $m\text{RATE}(t)$. In a subsequent chapter, we shall see how to express the speed-time function $m\text{RATE}(t)$ in terms of the mileage–time function $m(t)$. We might expect that this relation would ideally be of the form

$$s(t) = m\text{RATE}(t) \doteq (d/dt)m(t) \tag{1.8}$$

where the $\doteq$ sign means "is approximately equal to." In fact, we would find that the *numerical* relation expressed by Equation 1.8 would describe approximately the relation between s and m, provided that $s(t)$ is not changing very rapidly with t. The speedometer, which has as its input the angle through which the rear wheel has turned (and, hence, for a given tire diameter, how far the car has moved along the ground), thus functions as a physical *differentiator* to yield an output signal which is approximately equal to the *rate* at which the car is moving at any time. However, the speed reading at any instant is actually dependent on the speed of the car over a short time interval *prior* to that instant, whereas the derivative $(d/dt)m(t)$ in Equation 1.8 describes the purely mathematical concept of an "instantaneous velocity" at the instant t *only*. Hence, the value $s(t)$ and $(d/dt)m(t)$ will be similar only to the extent that the speed is not changing too rapidly.

A physical *differentiator* is thus a dynamic operator which yields an output signal that is *approximately* proportional to how much the input signal has changed during the small interval of time just past. It is a dynamic operator because the output value at any instant depends on the values of the input at prior instants of time. If we were to make a cartesian plot of the pairs of values of the speed $m\text{RATE}(t)$ and the mileage $m(t)$ for a number of different instants, we would not obtain points lying on a straight line. Even so, it would be a mistake to conclude that this operator is *not linear*. To see why this is so, let us define clearly and carefully what is meant by a *linear operator*.

Linear operators • We assume that it is physically meaningful to speak of "composing" any signal f from other signals, such as f_1 and f_2, where

$$f_1 + f_2 = f. \tag{1.9a}$$

Also, we assume that it is meaningful to speak of "multiplying" a signal by any scalar factor such as a. Then, we *define* an operator, L, to be *linear* if it is:

1. *Additive:* $$f\text{L} = (f_1 + f_2)\text{L} = f_1\text{L} + f_2\text{L}, \tag{1.9b}$$

and

2. *Homogeneous:* $$(af)\text{L} = a(f\text{L}). \tag{1.9c}$$

The *additive* property of a linear operator assures us that when a signal f is decomposable into the "sum" of other signals such as f_1 and f_2, the signal obtained by operating on f is likewise decomposable into the two signals, $f_1\text{L}$ and $f_2\text{L}$ that would result had this same operation been applied to f_1 and f_2 separately. Thus, the *principle of superposition* —that the effect of a cause, which is composed of component causes, is simply the "sum" of the effects of the component causes considered separately—may be applied to linear operators. This rather fuzzy statement is made precise by Equation 1.9b. The principle of superposition provides a fundamental approach to the study of linear systems since it often enables us to break a difficult problem into several simpler problems.

The *homogeneous* property assures us that a *linear operator commutes with any scalor.* That is, the signal fAL obtained by operating on f first with a *scalor* A is identical to the signal fLA obtained by operating first by L and then by A. [This easily follows from Equation 1.9c, since $(f\text{A})\text{L} = (af)\text{L}$ and $(f\text{L})\text{A} = a(f\text{L})$.] As a corollary of that property the result of operating on the *null signal* (i.e., the signal whose value at every instant is zero) is itself a null signal. Why?

To illustrate these ideas, let us consider again the simple linear system comprising a weight suspended from a spring, as shown in Figure 1.7. If the upper end of the spring is raised through a distance x, the weight would also move upward from its rest position through a distance y.

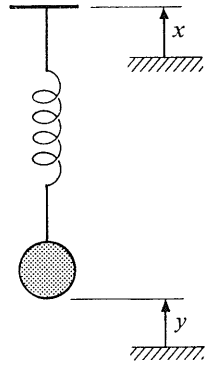
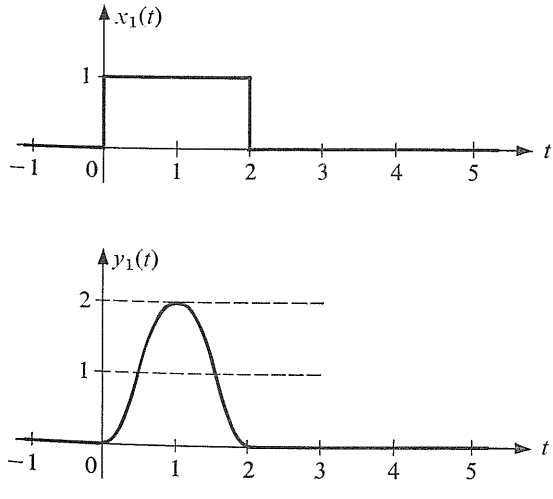
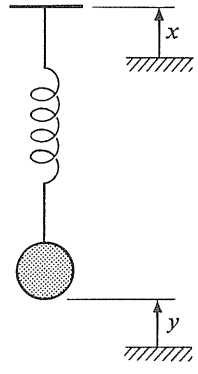


FIGURE 1.7 *A spring-weight system.*

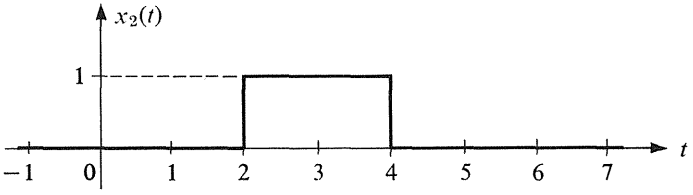
Prior to the instant $t = 0$, the system is at rest. (That is, $x(t) = 0$ and $y(t) = 0$ for all $t < 0$.) Subsequent to $t = 0$, the end of the spring is moved vertically so that at any instant t, the value $x(t)$ of the displacement is as described in the following question.

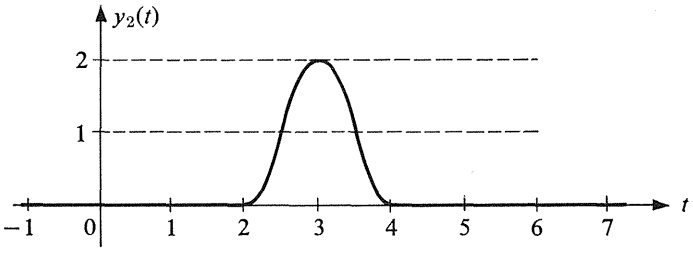
QUESTION 1.20 When $x_1(t)$ varies with time as shown in the upper plot of the figure, the corresponding displacement $y_1(t)$ is observed to vary with time, as shown in



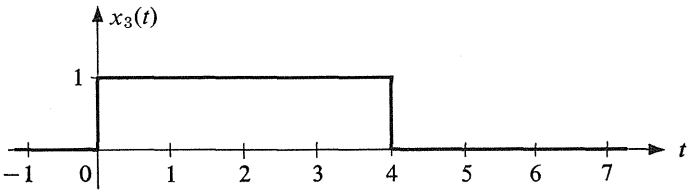


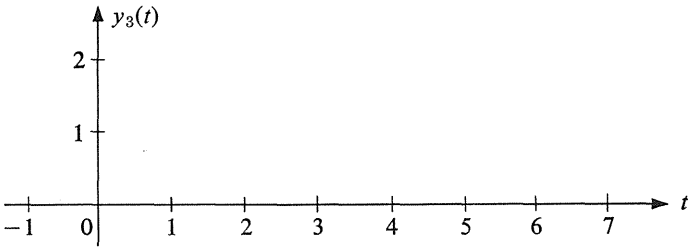
the lower plot. Furthermore, it is known that if the system had been left at rest until $t = 2$, and then the upper end raised so that $x(t) = 1$ for $2 \leq t < 4$, after which $x(t) = 0$, the response $y(t)$ would have varied with time as in the following graph. (We use sub-



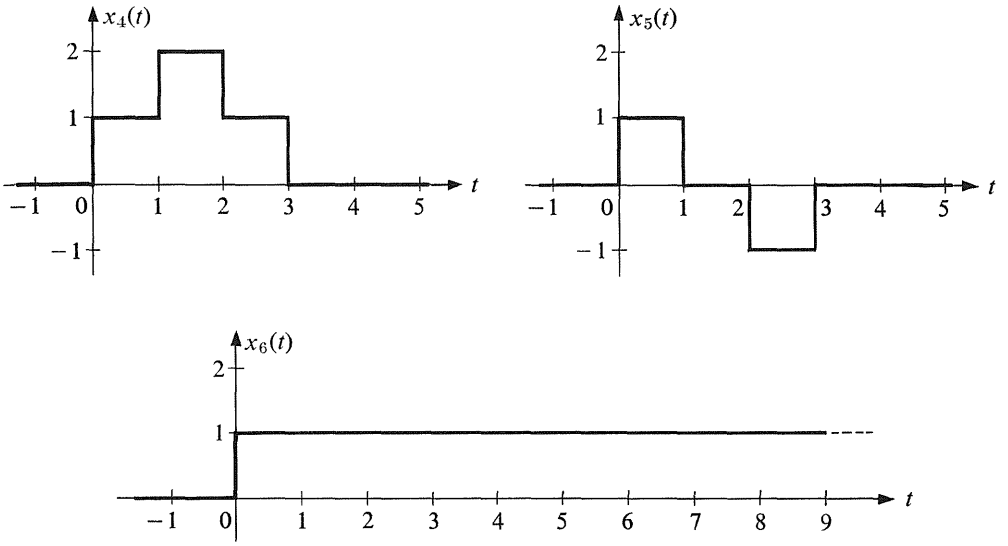


scripts "1" and "2" on the signals to distinguish the two cases described above.) If $x_3(t)$ were made to vary as sketched below, how would $y_3(t)$ vary as a function of time? [Show a graph of $y_3(t)$.] (Answer)





QUESTION 1.21 If x were to vary in accord with the functions $x_4(t)$, $x_5(t)$ and $x_6(t)$, what would be the corresponding response functions $y_4(t)$, $y_5(t)$, and $y_6(t)$? (Plot three graphs.) (Answer)



Delayors • In discussing Question 1.21, we referred to the operation of delaying a signal by some time interval, for instance, τ. Provided that τ is positive, this corresponds to a commonplace physical operation. Pending further notational refinements to be introduced in Chapter 3, we will call this operator a *delayor* and denote it by DELτ, where τ indicates the amount of delay.

To illustrate, the signal x_2 of Question 1.20 could be obtained from x_1 by delaying x_1 by two time units:

$$x_1\text{DEL2} = x_2$$

where

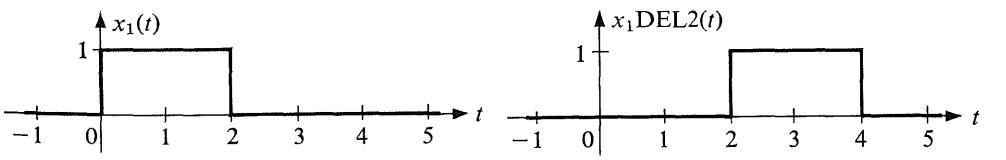
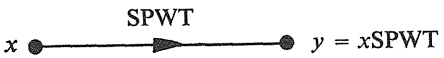


FIGURE 1.8 *The effect of a time delay of two units.*

In an earlier section, we observed that any *linear* operator commutes with any scalor. We now observe that any *stationary operator commutes with any delayor.* That is, if H is a *stationary* operator, the composite operator H DELτ is equivalent to DELτ H. Delaying the input to a stationary operator has the same overall effect as delaying the output. As a consequence, linear stationary operators enjoy a commutative property that allows us to manipulate them according to the familiar rules of algebra. When operators are *time-varying*, however, the order in which they are applied to a signal is significant, and care must be used in operator expressions. Most of the engineering systems with which we are concerned are linear and stationary, so in most of the subsequent development you will be able to manipulate operator symbols in accord with the

ANSWER TO QUESTION 1.20 The relation between the signals x and y is expressed by a spring–weight operator, say, SPWT:



and at any instant

$$y(t) = x\text{SPWT}(t).$$

We assume that SPWT is a linear operator.

Next, we observe that the signal x_3 is composed of x_1 and x_2:

$$x_3 = x_1 + x_2. \tag{1.10}$$

That this is so may be proved by observing that $x_3(t) = x_1(t) + x_2(t)$ for all values of t. But since SPWT is assumed to be a linear operator, we have by Equation 1.9b

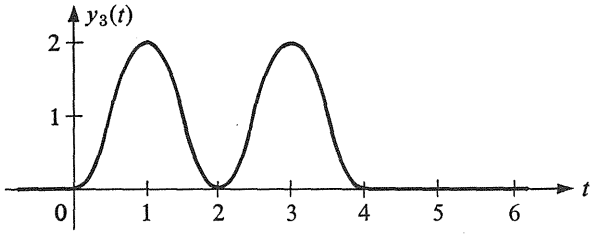
$$x_3\text{SPWT} = x_1\text{SPWT} + x_2\text{SPWT}$$

or

$$y_3 = y_1 + y_2.$$

Thus, at any instant t,

$$y_3(t) = y_1(t) + y_2(t). \tag{1.11}$$



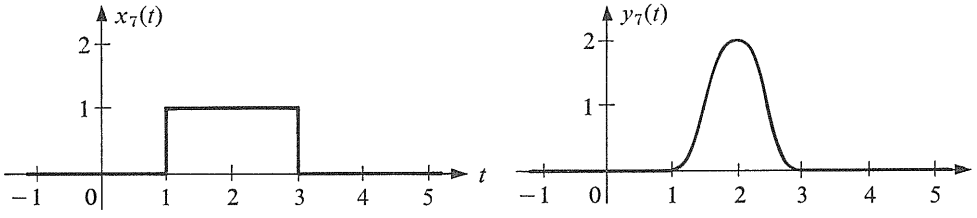
ANSWER TO QUESTION 1.21 To answer this question, you must have invoked another important property of SPWT that has not been explicitly mentioned. You no doubt observed in comparing $x_1(t)$ and $x_2(t)$ that x_2 is a time-delayed replica of x_1. That is,

$$x_2(t) = x_1(t - 2),$$

but the response y_2 to this *delayed input* is identical to the signal that would be obtained by delaying the original output y_1 by the same time interval. This means that the operation which SPWT performs on the input signal x does not depend on the absolute time at which x occurs. Hence, delaying the input signal by any amount produces the same overall result as delaying the original output signal by that same amount. Operators for which this is so are called *time-invariant* or *stationary* operators. The assumption that SPWT is a *stationary operator* is quite valid since it is likely that the physical

properties of the spring and of the weight are not changing with time. (But, can you describe circumstances in which the assumption of stationarity would not be valid even though the system is linear?)

If we make the plausible assumption that the system is stationary, then the response of the system to a signal which is x_1 delayed by one time unit will be identical to the response to x_1, delayed by the same time interval. Hence, we have the signals x_7 and y_7, which are plotted as follows.

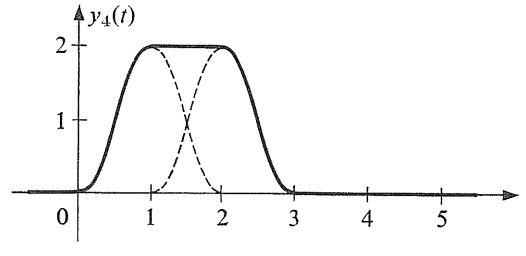


Now, evidently,

$$x_4 = x_1 + x_7.$$

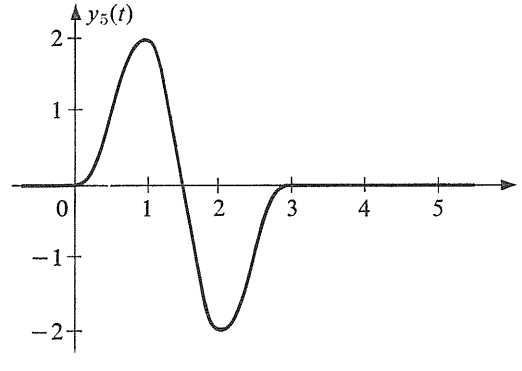
Hence,

$y_4 = y_1 + y_7.$



In a similar fashion,

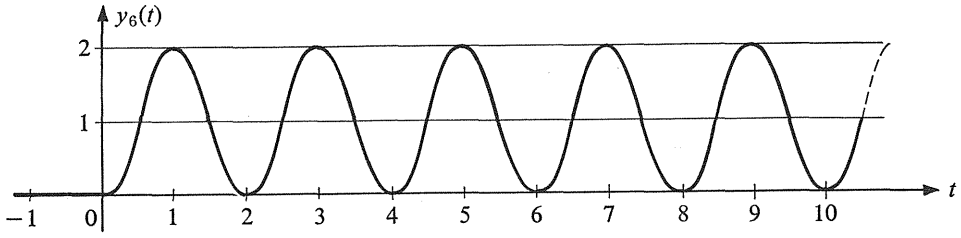
$x_5 = x_1 - x_7.$



Hence, $y_5 = y_1 - y_7$.

The response to x_6 is a "raised cosine" signal whose value at any instant, for $t \geq 0$, is

$$y(t) = 1 - \cos \pi t.$$



Similarly, $x_4 = x_1 + x_1\text{DEL1}$, where DEL1 is a unit delayor.

rules of ordinary algebra with which you are already familiar. To make these operator expressions more concrete, you may find it helpful to interpret all the operations in the remainder of this chapter and in Chapter 2 as being ordinary scalors which, as we have already seen, simply multiply the signal by a scalar factor.

Systems of Operators

Signal flow-graph symbolism may also be used to portray the simultaneous interactions among several signals of a system. It thus provides an easily visualized mathematical model for representing the relationships among the signals of a physical system. Figure 1.9 is a typical flow-graph involving five signals.

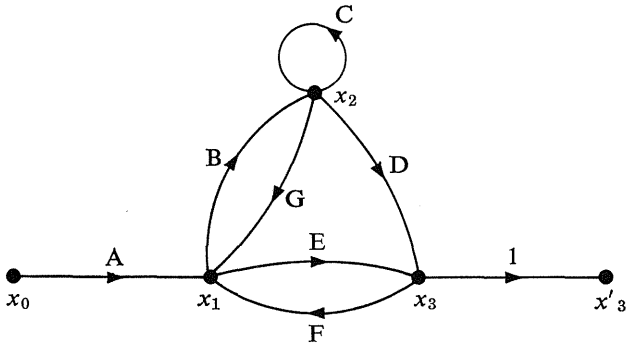


FIGURE 1.9 *A more complicated signal flow-graph.*

Why do we say this graph has five signals?

To interpret this graph, we must agree on the *syntax* or *rules* that govern these symbols and their manipulation.

We have already stated that each node represents a *signal.* For instance, one signal, x_k, is represented by *node k.* The *directed branch, jk,* may be thought of as outgoing from node j and incoming at node k. Alternatively, branch jk may be thought of as having its *input* at node j and its output at node k. The rules governing a flow-graph may then be stated as follows:

1. Signals travel over branches *only* in directions indicated by the arrows. (The branches are *not* wires, as in the ordinary electric circuit diagram, but represent operators.)
2. A signal in traveling over each branch is transformed by the operator of that branch. Thus, each branch represents an operator which acts upon the node signal at its input

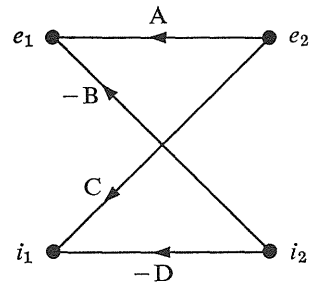
end and delivers the transformed signal to its output node. For instance, if the operator A in the graph shown above were a scalor, the signal x would be "multiplied" by the transmittance of A when flowing through the branch labeled A.

3. The signal at a node is composed of the algebraic sum of all signal flow *into* that node. Hence, at any instant t the value of a signal at each dependent node is the algebraic sum of the values of *all* signal flow *into* that node at that instant. (Observe that the node signal is in no way related to the number of branches which *leave* the node; only the signal flow *into* the node determines the node signal.)
4. The signal at each node serves as the input for all branches leaving that node.

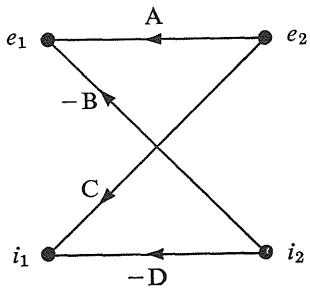
It is useful to define a *source node* as one with *no* incoming branches. A source node is used to represent a signal of *known* or specified value. All other nodes, having one or more incoming branches, are called *dependent* nodes. The signal at any dependent node is determined by rule 3. For other definitions, see the summary at the end of Chapter 2.

It is customary to designate the operator or transmittance of each branch by writing it alongside, and also to label the node signals, as has been done with *scalor* branches in the flow-graph of Question 1.22.

QUESTION 1.22 In accord with the definitions just given, select the *source* nodes of the graph. (Answer)



QUESTION 1.23 Assuming that the signals at these two source nodes are e_2 and i_2, respectively, find the total signal flow into the two *dependent* nodes and, hence, expressions for e_1 and i_1. Remember that a scalor branch delivers at its output end a signal whose value is simply the value of the input signal multiplied by the transmittance of the branch. (Answer)



Reversal of algebraic sign of signal • Of course, in the answers to Questions 1.22 and 1.23 there is no reason why the node could not represent $-i_2$ instead of $+i_2$. In setting up a flow-graph for solution on an analog computer, it is often desired to change the sign of the signal at some node. The question then arises as to what should be done to the transmittances of the branches which are entering and are leaving that node so that the resulting graph will be correct. That is, starting with the flow-graph at the left in Figure

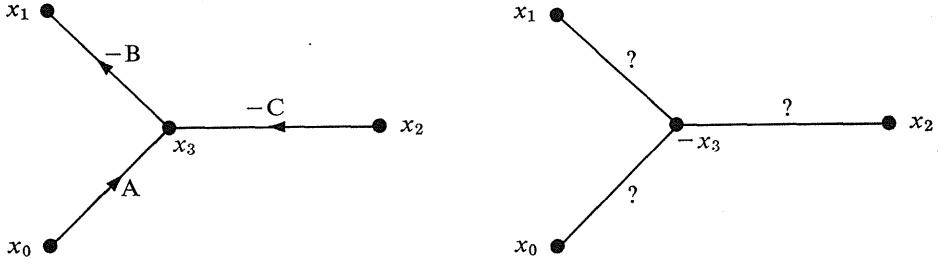


FIGURE 1.10 *What is the effect of replacing x_3 by $-x_3$?*

1.10, we wish to find a modified graph which is equivalent to the first graph, except that the center node now represents $-x_3$ instead of x_3 as in the original graph.

The physical meaning of "changing the sign" of a signal should be kept in mind. Remember, signals have an additive property so that an expression like $x_1 + x_2 = x_3$ corresponds to some meaningful physical composition of observables. For instance, x_1 and x_2 might be two successive *upward* displacements at the end of the spring in Question 1.20. The *total* upward displacement x_3 is composed of the "sum" of these two displacements. However, suppose that the second displacement x_2 were such that the total displacement x_3 is *null*. That is $x_1 + x_2 = 0$ (where the value $0(t)$ of the null signal is zero at every instant). The particular signal x_2 which when combined with x_1 will yield the null signal is denoted by $-x_1$. The *value* of $-x_1$ at any instant is the negative of the value of x_1 at the same instant:

$$\underline{-x_1}(t) = -\underline{x_1(t)} \tag{1.12}$$

Also, remember that although the branches of a flow-graph look like a wire in an electric

ANSWER TO QUESTION 1.22 The source nodes represent the signals e_2 and i_2. Neither of these nodes has any incoming branches. Hence, the signals e_2 and i_2 must be dictated solely by outside *proclamation* rather than through any dependence upon the signals at the other nodes (as must all *dependent*, i.e., nonsource, nodes). Unless the signal at a source node is specified, we shall, by convention, assume *that its value is zero at every instant of time.*

ANSWER TO QUESTION 1.23

$$e_1 = e_2\mathrm{A} - i_2\mathrm{B},$$

$$i_1 = e_2\mathrm{C} - i_2\mathrm{D}.$$

circuit, the *rules* governing signals in a flow-graph are quite different from those for an electric circuit. Before answering the next question, note carefully that Figure 1.11

FIGURE 1.11

means $x\mathrm{A} = y$, whereas Figure 1.12 means $x = -y\mathrm{A}$.

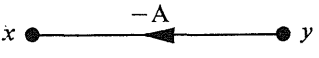
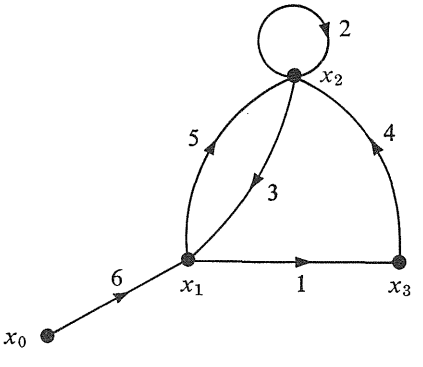


FIGURE 1.12

QUESTION 1.24 Write below one or two sentences to convey the rules needed to correctly assign the proper transmittances to the modified graph referred to above in such a way that the values of x_1, x_0, and x_2 are the *same* in the two graphs, and the value of $-x_3$ will be the negative of x_3. (Answer)

QUESTION 1.25 It is desired to modify the graph shown below so that one of the nodes represents $-x_2$ instead of x_2. Draw a modified graph having the requisite changes in the branch transmittances. Note particularly any special property of the loop on node x_2, and consider its consequences before answering. (Answer)



The particular branch referred to in the answer to Question 1.25 *enters* the node and it *also leaves* it! Hence, the algebraic sign of its transmittance was reversed *twice*, whereas the signs of all other branches were reversed only *once*. Before continuing with your reading, verify that the two graphs considered in Question 1.25 are equivalent to the two sets of equations:

x_0 is a source	x_0 is a source
$x_1 = 6x_0 + 3x_2$	$x_1 = 6x_0 - 3(-x_2)$
$x_2 = 5x_1 + 2x_2 + 4x_3$	$-x_2 = -5x_1 + 2(-x_2) - 4x_3$
$x_3 = x_1$	$x_3 = x_1$

It is important that you clearly understand at this point how these equations may be written, once the corresponding flow graph is given. They are obtained by writing the total signal flow *into* each node. Note particularly that the equation for x_2 contains x_2 also on the right-hand side. That is, the value of x_2 depends in some measure on its own value. This indicates the presence of *feedback* in the signal relationships and, as we shall soon see, it is associated with the existence of closed paths or *loops* in the flow-graph.

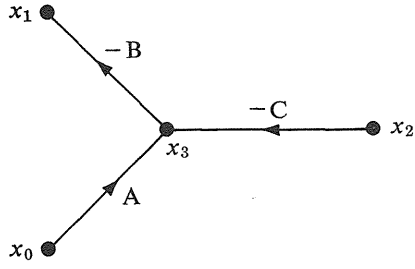
More definitions • It will be helpful to define a few terms needed in the sequel. A *path* is *any* continuous succession of branches traversed in the indicated branch directions. You may think of the graph as a road map of one-way streets. Any route through the graph is permissible, provided that the one-way directions are observed. It *is* permissible to *revisit* any node and to traverse the same branch *any number of times.* A path may be uniquely described or labeled by listing the labels of the branches in the order that they are traversed. For instance, in Figure 1.13, one possible path from the *source* node, x_1, to node x_6 is ADGEC, as shown at the right.

Graphs in which there are only a finite number of paths are sometimes called *cascade* (or *open*) *graphs* because it is a relatively simple matter to express each signal along a path in terms of the previous signal. Thus, all dependent signals are ultimately expressible in terms of the source signals by a cascade of substitutions.

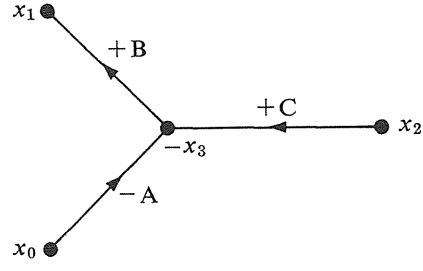
For instance, starting with the source signal x_1 as given, we may express all subsequent signals encountered along these paths in terms of signals previously evaluated. Thus, showing the evaluation of each signal alongside the node, we have Figure 1.14.

QUESTION 1.26 How many different paths from x_1 to x_6 exist through the graph shown in Figure 1.13? Describe each path by writing its label. (Answer)

ANSWER TO QUESTION 1.24 An easy way to write this is to say: "Change the algebraic sign of the transmittances on all branches that *enter* x_3 and *also* on all branches that *leave* x_3." If this rule is used, we have



(Original graph)



(Modified graph)

If the sign of the node signal is changed, one merely changes the algebraic sign of the transmittances of all the branches entering that node, and also of all branches leaving that node.

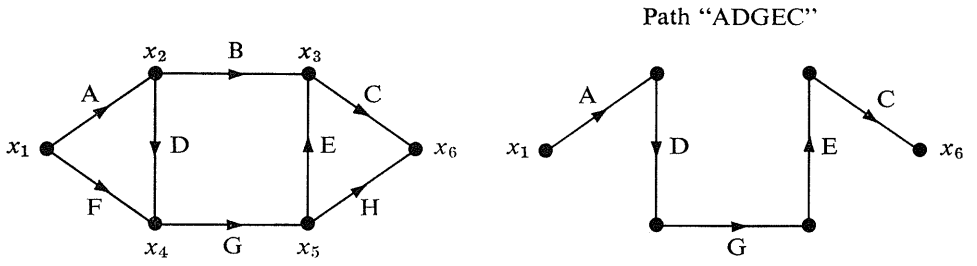


FIGURE 1.13

Two paths are *different* if they differ in any respect. For instance, the path, ADGE, from x_1 to x_3 differs from (although it is a segment of) the path ADGEC from x_1 to x_6.

QUESTION 1.27 Do you observe any noteworthy relationship between this expression for x_6 in terms of x_1 and the products of the branch transmittances, and the answer to Question 1.26? (Answer)

This result illustrates the fact that the signal flow from one node to another occurs over *all possible paths* between the two nodes. (The fact that the operators are all *linear* assures us that there is no interference among the signals as they pass through the branches.)

To proceed, we must distinguish several kinds of paths. An *open path* is a path along which no node appears more than once. A *closed path* begins and ends on the *same* node. In the graph just considered, *all* paths were *open*, so there were *no* closed paths. However, in the graph shown in Question 1.28, there are both *open* and *closed* paths.

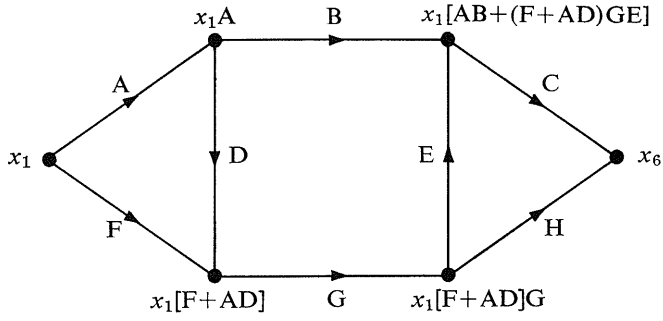


FIGURE 1.14

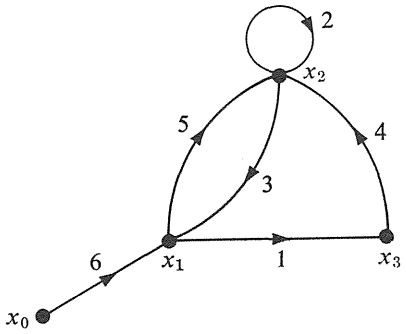
Whereupon we may express x_6 as

$$x_1[AB + (F + AD)GE]C + x_1[F + AD]GH = x_6.$$

Collecting terms, we have

$$x_1[ABC + FGEC + ADGEC + FGH + ADGH] = x_6.$$

QUESTION 1.28 How many different *open* paths exist from x_0 to x_2 in the graph? Show them in a sketch. (Answer)

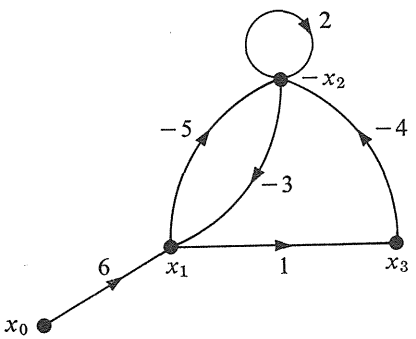


Unlike the previous graph, in which there were a finite number of paths, the graph considered in Question 1.28 has an *infinite number* of paths because any of several closed paths may be traversed an unlimited number of times. To describe this possibility, we next define a special type of closed path.

A closed path is said to be a *loop* if each node along the path is traversed *only once* in making a complete circuit.

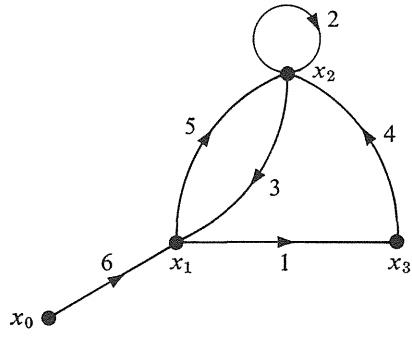
Using this definition of loop, we see that a *cascade graph* has no loops. Conversely, in a *loop graph*, *each* branch of the graph forms part of at least one loop.

ANSWER TO QUESTION 1.25 Your graph should look like this one. Note that the branch labeled 2 does *not* have its transmittance altered. (Or, more correctly, it has its algebraic sign changed *twice*, so the effect is as if no alteration were made.)



ANSWER TO QUESTION 1.26 There are five different paths through the graphs, ADGEC, ADGH, ABC, FGH, FGEC.

QUESTION 1.29 To make certain that you understand these definitions, examine the accompanying graph and tell how many *loops* there are. Also, tell how many *open paths* there are from x_2 to x_3. Finally, state the number of paths from x_1 to x_3. Remember, in traversing a loop you can pass through any node *only once*; also, two loops must differ by at least one branch (i.e., the *same* loop may be traversed by starting at any point along the loop). (Answer)



QUESTION 1.30 Is the graph shown in Question 1.29 a loop graph? Is it a cascade graph? (Answer)

Flow-Graph Representation of an Electrical System

In this section, we shall represent the signal relationships in an electrical system comprised of *resistors* connected to a *battery* in the configuration shown in Figure 1.15. The elements, or physical building blocks, from which this system is constructed are of only two kinds—resistors and a signal source (i.e., battery). Figure 1.15 shows the *things* composing the electrical system. It also shows the *signals* (i.e., the *voltages* and *currents*) which may be observed in this circuit by means of suitable measuring instruments.

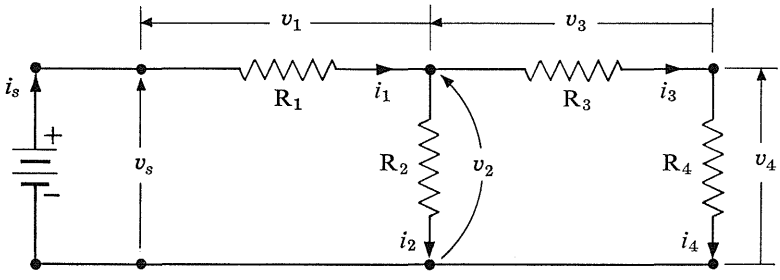


FIGURE 1.15 *An electrical system.*

ANSWER TO QUESTION 1.27 The signal components composing x_6 constitute the sum total of the "flows" of x_1 over each and every path from x_1 to x_6.

Voltage and current are two very important physical observables associated with any electrical system. (There are other observables, too, such as charge and flux linkage.) We need not concern ourselves with the physics except to note that *current* is the flow of electric charge (which behaves like an incompressible fluid) and that *voltage* expresses the work required to move electric charge from one point to another.

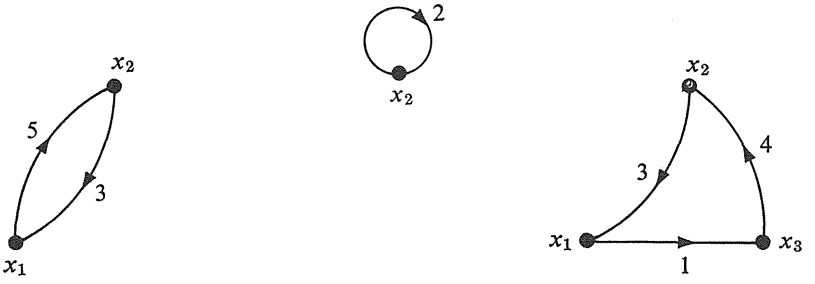
An ammeter may be connected in *series* with a conductor to measure the current through the conductor. Since there are two terminals on an ammeter, it may be connected in the circuit in either of two ways. To avoid ambiguity, one of the terminals is marked (often with a plus sign) and a *positive* reading of the ammeter then signifies that the current is entering the ammeter by the marked terminal and is leaving by the other. A negative reading signifies that the current is entering the *unmarked* terminal and leaving by the marked terminal.

It is often convenient to classify signals into two types—*through* and *across* observables —depending on how they are measured. Current is a through-type observable because an ammeter is connected so that current being measured passes *through* its terminals.

A *voltmeter* may be connected between two points in a circuit to measure the *work* or *effort* that would be required to move a unit charge from one point to the other. Again, there are two terminals on a voltmeter and two ways in which it may be con-

ANSWER TO QUESTION 1.28 There are *two* open paths from x_0 to x_2: (6, 5) and (6, 1, 4).

ANSWER TO QUESTION 1.29 In this graph there are *three loops*:



There is only one *open* path from x_2 to x_3. There are an infinite number of *paths* from x_1 to x_3. The "infinite number of paths" may be a little confusing at first. But, one merely goes around any of the loops 1, 2, 3, . . . , number of times before entering the x_3 node. For instance, the path $x_1\, x_2\, x_2\, x_2\, x_1\, x_2\, x_2\, x_2\, x_2\, x_1\, x_3$, starts at x_1 and terminates at x_3. (Note, however, that the path indicated is by no means an *open path*, since two of the nodes are encountered more than once.)

ANSWER TO QUESTION 1.30 It is *not* a loop graph because there is one branch, 6, that is not part of a loop. It is *not* a cascade graph because the graph contains loops.

nected in the circuit. To avoid ambiguity, one of the terminals is marked (often with a plus sign) and a positive reading of the voltmeter signifies that electric charge in moving through the circuit from a point connected to the + terminal to a point connected to the unmarked terminal will *do* work *on* the circuit. Voltage is an across-type observable because a voltmeter is connected so that the voltage being measured appears *across* its terminals.

Many important kinds of electric circuit elements have only two wires or terminals at which they may be connected to other circuit elements. At these two terminals, we may measure the voltage and the current by connecting a voltmeter and an ammeter in the circuit as shown in Figure 1.16. The signals i and v, whose values at any instant may be, in principle, measured by the ammeter and voltmeter of Figure 1.16(a) are commonly designated directly on the circuit schematic, as in Figure 1.16(b).

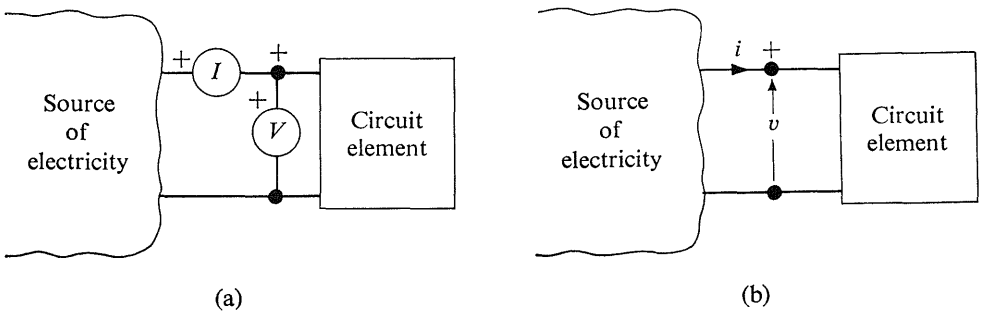


FIGURE 1.16 *The ammeter and voltmeter, shown in (a), may be used to measure the values of the current i and voltage v shown in (b). The arrows in (b) show how the meters are to be connected into the circuit.*

The current through a conductor may be designated by drawing an arrow on the conductor and writing the symbol for the current alongside the arrow. The direction of the arrow specifies how the ammeter is to be connected into the circuit (i.e., so that the arrow specifies flow into the marked terminal). The direction of the arrow implies *nothing whatsoever* about the physical movement of electrical charge through the wire—it merely specifies how a measuring instrument might be connected so that we may assign unambiguously a *value* to whatever motion of charge may occur.

Similarly, the voltage between two terminals may be designated by drawing a light line from one terminal to the other with an arrow at one end with a gap near its middle for the symbol denoting this particular voltage. The direction of the arrow specifies how the voltmeter is to be connected across the circuit (i.e., so that the arrowhead points toward the marked terminal). The direction of the arrow implies *nothing whatsoever* about the actual effort required to move charge from one terminal to the other—it merely specifies how a measuring instrument might be connected so that we may assign unambiguously a *value* to whatever effort may be required.

QUESTION 1.31 Redraw the circuit of Figure 1.15, showing explicitly how you would connect ammeters and voltmeters into the circuit so as to measure each of the 10 signals indicated thereon. (Answer)

Any number of attributes may be observed about a resistor—its size, shape, color, weight, and smell, to name a few—yet none of these are of any relevance in so far as the *electrical properties* of the circuit are concerned. The basic electrical property that describes a resistor is its *resistance.* The resistance, as we have already remarked, is an abstract designation for that particular *physical* property of a resistor that relates the *voltage across* the *terminals* of the resistor to the *current through* the resistor. The resistance symbol denotes either of the flow-graph relations at the right in Figure 1.17.

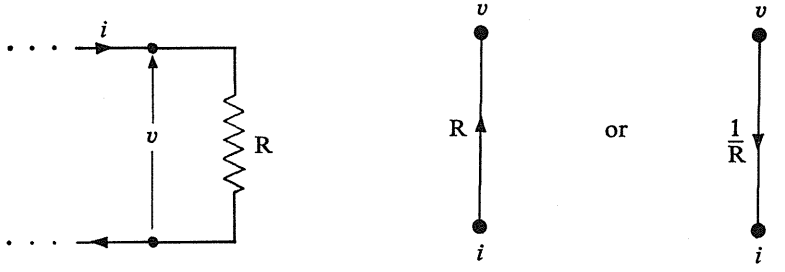
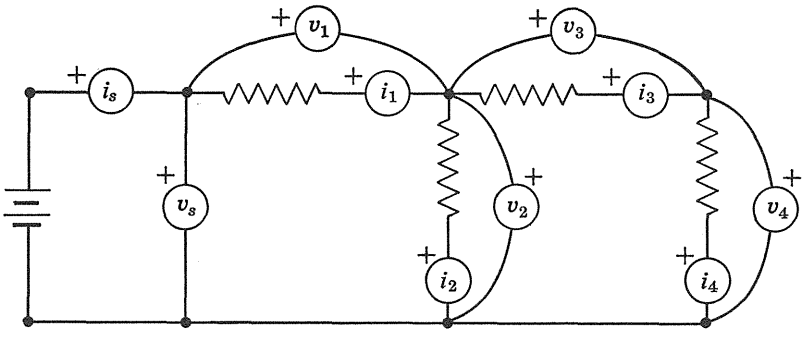


FIGURE 1.17 *Left, a circuit-diagram symbol for a resistor; right, the equivalent signal flow-graphs.*

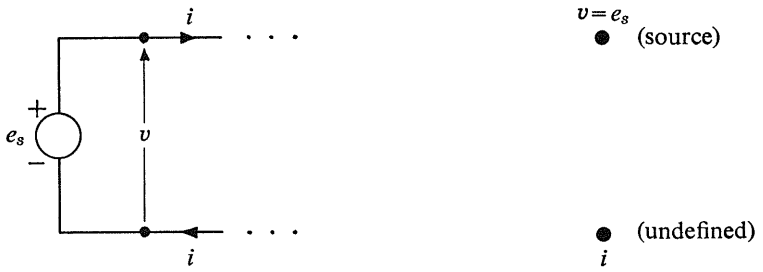
The value of the resistance R may be expressed by forming the ratio of the value of the voltage across the terminals at any instant to the value of the current at that same instant. Resistance is thus the transmittance of a scalor that operates on a current to yield a voltage. In an ideal resistance, this transmittance is a constant that is independent of the value of the current or voltage. (This is not strictly so in *physical* resistors, where the resistance depends on such things as temperature, magnetic field in the vicinity, etc.)

Also forming part of this system is a *signal source*—in this case a battery. The important electrical property of the battery is the voltage that it delivers *across* its terminals. For many batteries, this voltage has a value that is nearly independent of the value of the current *through* it. We may abstract this important property by defining an *ideal voltage source* which delivers a voltage of specified value *regardless* of the value of the current.

ANSWER TO QUESTION 1.31



In subsequent discussion of this system, we shall replace the actual physical elements by idealized *resistances* and an *ideal voltage source*. Each of these elements has two terminals at which it connects to other elements. It is completely described by specification of the relation between the current *through*, and voltage *across*, these terminals.



Ideal voltage source element

Flow graph representation

FIGURE 1.18

In describing the current and voltage associated with a circuit element, it is customary to choose the connections of the ammeter so that both the current and voltage arrows point toward the same terminal. Ammeters and voltmeters have been calibrated so that the product of the indicated values of the current and voltage yields the electric power *into* the circuit element. For a resistor, this power must never be negative. Accordingly at any instant these current and voltage values must have the *same* algebraic sign (thus making their product *nonnegative*). By this convention, assigning the direction of the current arrow automatically determines the direction of the volt arrow, and *vice versa*.

A physical resistor comes about as close to being an ideal scalor as any device one can think of. For a good resistor, the voltage v is directly proportional to the current i even when the current is varying extremely rapidly from one instant to the next. The fact that the electrical system of Figure 1.15 is known to be constructed from four resistors, therefore implies that four of the current and four of the voltage signals are related as Figure 1.19.

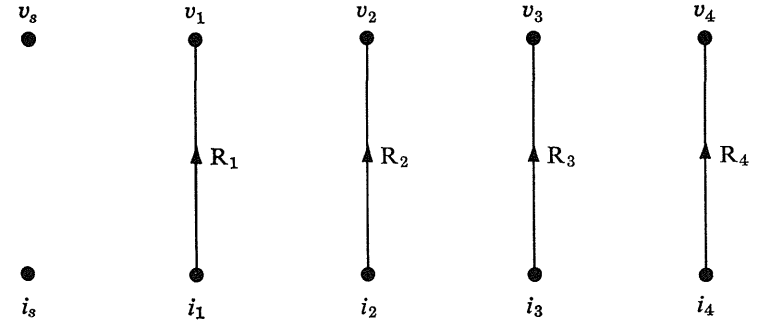


FIGURE 1.19 *Flow-graph representation of several current-voltage relations.*

Besides the resistance relations between current and voltage, there is a set of relations among *only* the currents, and still another set of relations among *only* the voltages. These two sets of relations are described by "Kirchhoff's *current law*" and "Kirchhoff's *voltage law*." They exist because current and voltage are *additive observables*.

You may think of an electric current as the flow of an *incompressible* fluid. At a junction of three conductors, the sum of the three values of the current flow measured toward the junction must add to zero at any instant, that is,

$$i_1 + i_2 + i_3 = 0.$$

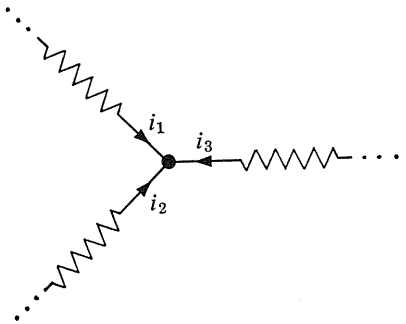


FIGURE 1.20 *A physical junction of three conductors carrying currents* i_1, i_2, *and* i_3.

This relation holds for any number of conductors, and is expressed by the general equation

$$\sum_k i_k = 0. \tag{1.13}$$

Kirchhoff's current law describes a condition that must be satisfied by the currents flowing into any junction of a circuit. It is a simple matter to rearrange the terms in the equation to define any one of the currents in terms of the remaining currents. For instance, for the above example, we may express i_3 in terms of i_1 and i_2 as $-i_1 - i_2 = i_3$. Hence, the fact that the currents into this junction must sum to zero may be written in flow-graph symbolism as

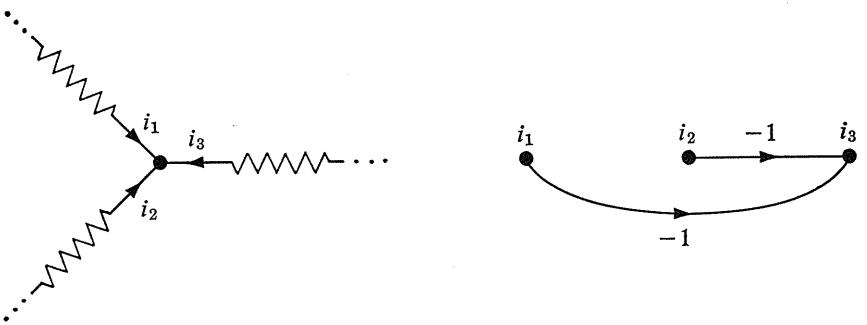


FIGURE 1.21 *The current* i_3 *depends on the current* i_1 *and* i_2.

In this graph, i_3 is represented as a *dependent* signal and i_1 and i_2 are treated as *independent* (or source) signals.

QUESTION 1.32 What other two signal flow-graphs are also valid descriptions of the relation between i_1, i_2, and i_3 as imposed by Kirchhoff's current law? (Hint: in the example just considered, i_3 was dependent on i_1 and i_2. Can you find a graph in which i_2 is dependent on i_1 and i_3? And i_1 on i_2 and i_3?) (Answer)

The preceding question reveals that we may think of any two currents at the junction as being the "cause" and the third as the "effect." Which of these three possibilities we use depends on how these three currents are related to the other signals in the circuit. We have already remarked that there is a good deal of arbitrary choice available here. We shall illustrate by continuing with the examination of this circuit.

Suppose that we wish to use Kirchhoff's current law to express the relations between all the currents in the circuit, but treating i_2 and i_4 as independent signals (i.e., as sources).

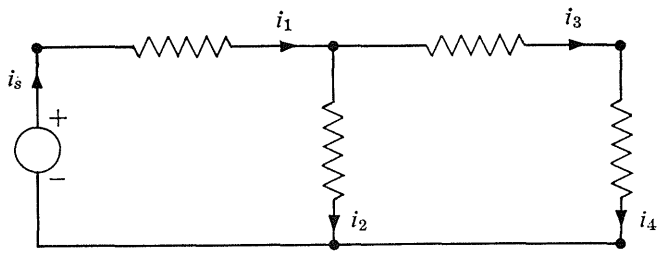
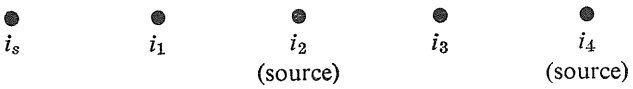


FIGURE 1.22 *Current flow through a given circuit.*

QUESTION 1.33 Treating i_2 and i_4 as *source* nodes and all other current nodes as *dependent* nodes, draw the branches between the nodes shown below to show the proper relations. (Answer)



Next, suppose that we adjust the battery voltage to establish some specified value for the current i_4. We would now like to express the *signals* throughout the circuit *in terms of* i_4. We now start with the relationships between the currents found in Question 1.33, in which i_2 and i_4 are *independent* signals, and see if we can express some of the voltages in terms of i_4.

Initially, our flow-graph looks like Figure 1.23. But the current–voltage relations for

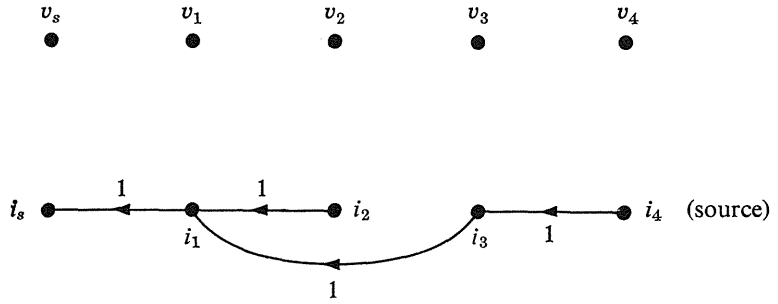


FIGURE 1.23 *All currents can be expressed in terms of i_2 and i_4.*

the resistances R_4 and R_3 state that $i_4R_4 = v_4$ and $i_3R_3 = v_3$. Hence, we can immediately add two more scalor branches to the flow-graph, as in Figure 1.24.

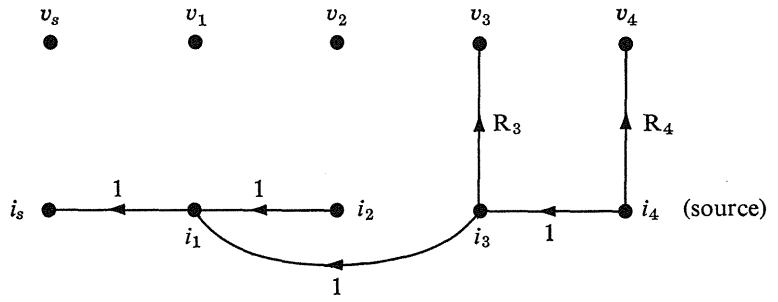
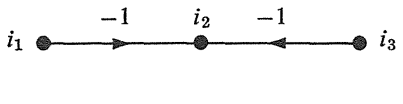


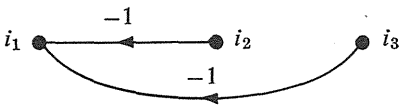
FIGURE 1.24 *In a resistor voltage is proportional to current.*

Now we come to an impasse. In the physical circuit only i_4 has been specified. From physical considerations, this should be sufficient to determine not only the battery voltage necessary to produce the specified value of i_4 but *all* of the other currents and voltages throughout the circuit as well. Hence, we would expect that i_2 must be expressible in terms of the other signals in the circuit. In particular, if we know v_2, we can certainly find i_2 because they are related by $v_2R_2^{-1} = i_2$. But how are we to find v_2 ?

ANSWER TO QUESTION 1.32 Each of the three currents may be expressed as an additive combination of the other two. Thus, equally valid descriptions are



or $-i_1 - i_3 = i_2$



or $-i_2 - i_3 = i_1$

At this point, Kirchhoff comes to the rescue with his *voltage law*, which asserts that *the total change in voltage around any closed circuit is zero.*

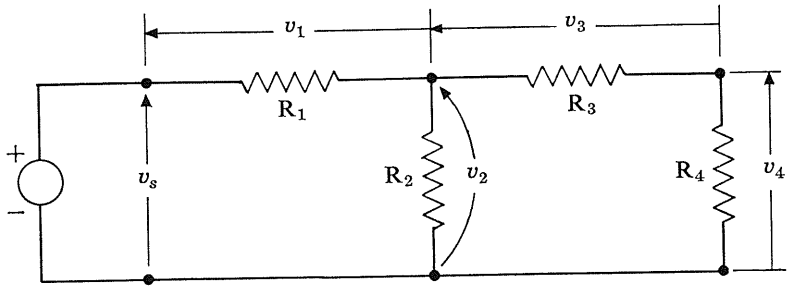


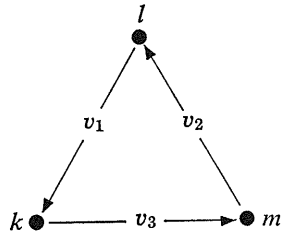
FIGURE 1.25 *The voltage across each resistor.*

Thus, starting at the bottom end of R_2 and tracing a closed path through R_3 and R_4 back to the bottom end of R_2, we find that the total change is given by the *sum of the voltages* across each of these three resistances:

$$v_2 + (-v_3) + (-v_4) = 0.$$

The *minus signs* are needed because v_3 and v_4 are defined to be the voltages measured between the terminals in the *opposite* direction than that in which we are traversing the loop.

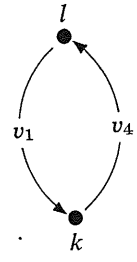
More generally, consider *any* three terminals, indexed by k, l, and m, between which appear the voltages v_1, v_2, and v_3 as indicated in Figure 1.26. Kirchhoff's voltage law asserts that



$$v_1 + v_2 + v_3 = 0.$$

FIGURE 1.26

Furthermore, by applying this same law to the two voltages v_1 and v_4 associated with any pair of terminals, such as those shown in Figure 1.27, it follows that



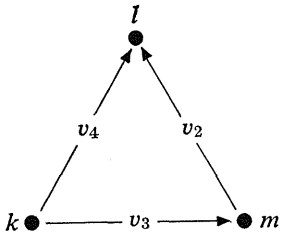
$$v_1 + v_4 = 0$$

FIGURE 1.27

or that

$$v_1 = -v_4$$

Hence, we may alternatively express the voltage relation among the three terminals considered previously as



$v_4 = v_2 + v_3$

FIGURE 1.28

in Figure 1.28. This *additive* law for the composition of voltages is exactly like that of geographical elevation. Thus, if point m is v_3 "higher" in voltage than point k, and point l is v_2 "higher" in voltage than point m, then point l is higher in voltage than k by the amount $v_4 = v_2 + v_3$. Note particularly that the numerical *value* of voltage may be either *positive* or *negative*. The arrow used to designate the voltage on the circuit merely establishes the *sense* of the measurements and in *no way* implies that the voltage has a positive or a negative *value*.

In our electric circuit, Kirchhoff's voltage law demands that $v_3 + v_4 = v_2$. This relation, together with the relations that $v_2 R_2^{-1} = i_2$ and $i_1 R_1 = v_1$, put into our flow-graph give:

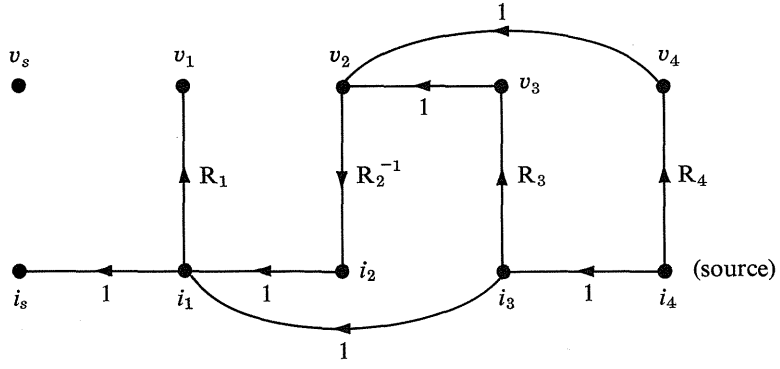
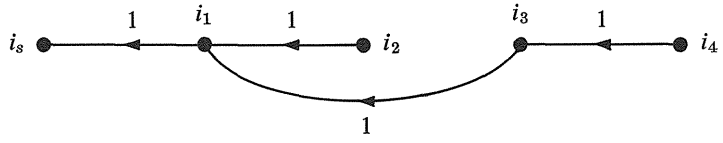
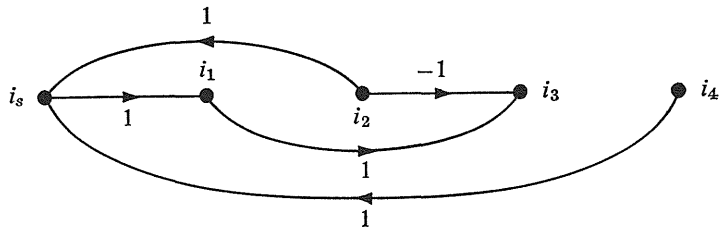


FIGURE 1.29 *The voltage v_2 depends on the voltages v_3 and v_4.*

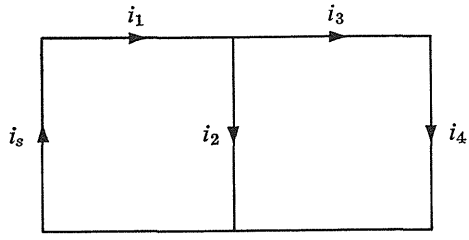
ANSWER TO QUESTION 1.33 Your graph may look like



or like



or like any of several other possibilities. In each case, however, we note that every current can be expressed in terms of i_2 and i_4. These two currents are called a *current basis* because every current in the circuit may be expressed as some linear combination of them. It should be evident from this picture of current flow that



$$i_3 = i_4,$$
$$i_1 = i_2 + i_3,$$
$$i_s = i_1.$$

The first flow-graph simply restates these relationships in different notation. What are the corresponding sets of equations for the second flow-graph?

QUESTION 1.34 How else can you use Kirchhoff's voltage law to express the battery voltage v_s (the only remaining signal not already expressed in terms of i_4), in terms of the other signals? Draw the necessary branches on the graph above, showing their correct transmittances. (Answer)

We shall now show how we may evaluate any of the signals at the dependent nodes in terms of the specified value of i_4 and the values of the resistances. First, we note that this is a *cascade* graph since it does *not* contain any closed loops. All of the nodes in this graph are *dependent* except for node i_4, which is a *source*. The v_s and i_s nodes are *sinks*, since they have no outgoing branches. In an open graph, once the value of the source-node signal has been specified, the values of the signals at all other nodes are uniquely determined by the graph. For instance, the value of v_2 is:

$$v_2 = i_4R_4 + i_4R_3 = i_4(R_4 + R_3).$$

Here, the first and second terms are associated with the following two open paths:

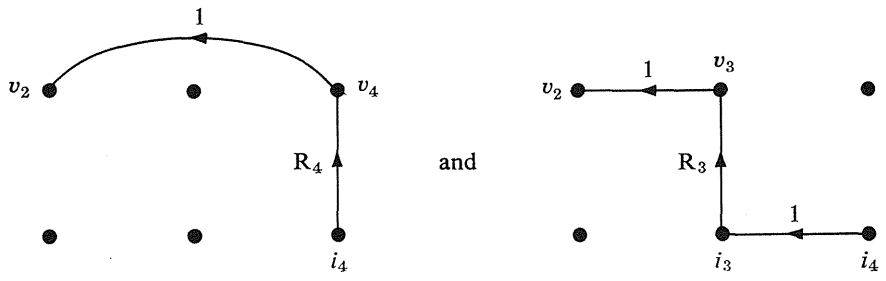


FIGURE 1.30

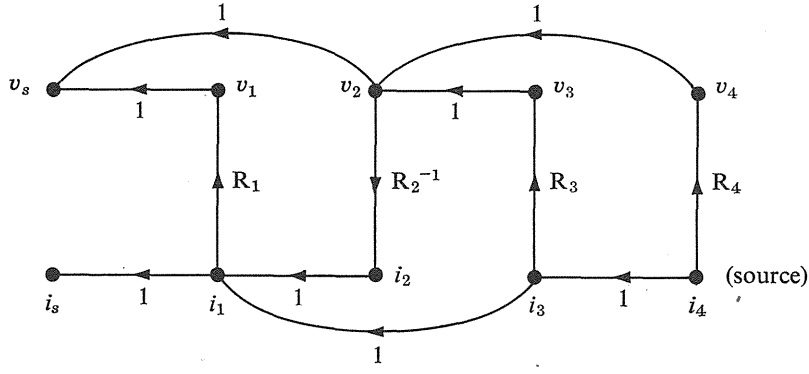
The signal contributed by any open path depends partly on the *path transmittance*, which is the product of the branch transmittances in that path. The path transmittance of the first open path is $1 \cdot R_4 = R_4$, and that of the second path is $1 \cdot R_3 \cdot 1 = R_3$.

The signal delivered at the output of an open path may then be expressed as the path transmittance multiplied by the signal at the input of the open path. It is easy to see why

ANSWER TO QUESTION 1.34 The battery voltage v_s can be expressed as the sum of the voltages across R_2 and R_1:

$$v_s = v_1 + v_2.$$

Hence, the complete graph, in which every signal is expressed in terms of i_4, becomes



Notice that by glancing at the flow-graph you can tell instantly that all other signals are expressed in terms of i_4. This is the only node that has no incoming branches. You may also note that there is no branch directly connecting i_s to v_s. This is because both v_s and i_s are determined, once i_4 has been specified. Any source that would *simultaneously* produce these values of current and voltage at its terminals would be consistent with the specified signal i_4.

this must be so generally. Consider the open path in Figure 1.31. Clearly, $x_1 = x_0\mathrm{A}$. But $x_2 = x_1\mathrm{B} = x_0\mathrm{A}\cdot\mathrm{B}$. Hence, the transmittance from x_0 to x_2 is simple $A \cdot B$. In

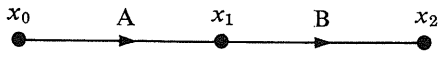


FIGURE 1.31

fact, if we were not interested in the value of x_1, the open path may be replaced by a single equivalent branch, as in Figure 1.32.

FIGURE 1.32 *The operators* A *and* B *of Figure* 1.31 *may be combined into a single operator* AB.

QUESTION 1.35 For the complete flow-graph shown in the answer to Question 1.34, determine the number of different open paths between i_4 and i_1. Draw these paths and indicate the path transmittance of each. What is the total transmittance from i_4 to i_1? (Answer)

QUESTION 1.36 By continuing the reduction of the graph in the manner illustrated, find the graph transmittance from i_4 to v_s, and also the graph transmittance from i_4 to i_s. Show these two graph transmittances as two branches in an equivalent reduced flow-graph having only three nodes (i_4, v_s, and i_s). (Answer)

The original graph has been correctly reduced to only two branches, one giving the value of v_s in terms of the assumed value i_4, and the other giving the value of i_s in terms of i_4. The fact that all the other nodes have been eliminated corresponds to the algebraic process of eliminating the corresponding v's and i's from the original set of equations.

Starting from the original ladder circuit, we could have used the principles of circuit theory to write a set of simultaneous algebraic equations. For the particular circuit shown, ten equations could be written. It is not impossible to solve ten equations in ten unknowns, but it is tedious. The flow-graph offers an alternative method of solution, and the graph provides a picture of the entire set of ten equations. The relations among a set of variables can be expressed *either* as a set of simultaneous equations *or* graphically in a flow-graph. You already know how to solve sets of equations. The flow-graph methods you are now learning gives you an alternate approach which for some purposes will be more useful.

Just as one "solves" a set of equations, so we shall "solve" a flow-graph. Corresponding to every manipulation of the equations, we shall establish a corresponding manipulation of the graph. The important feature of the flow-graph is that it provides a single "picture" of the *entire set* of equations, which is often helpful in clarifying the relationships among the various signals.

The open (or cascade) graph that was developed for this ladder circuit is particularly easy to solve. One simply starts with the value i_4 and then expresses each of the other

signals in terms of i_4. Note that it is unnecessary to solve *simultaneously* any equations. In fact, this will always be the case with *open graphs.* If, however, the flow-graph contains loops, then the simultaneous solution of the corresponding set of equations will be required.

Just as we can write several different sets of equations to represent a *given* circuit, so also we may formulate a variety of different flow-graphs. For instance, shown in Figure 1.33 is a different flow-graph formulation of the signals in this *same* ladder circuit. This

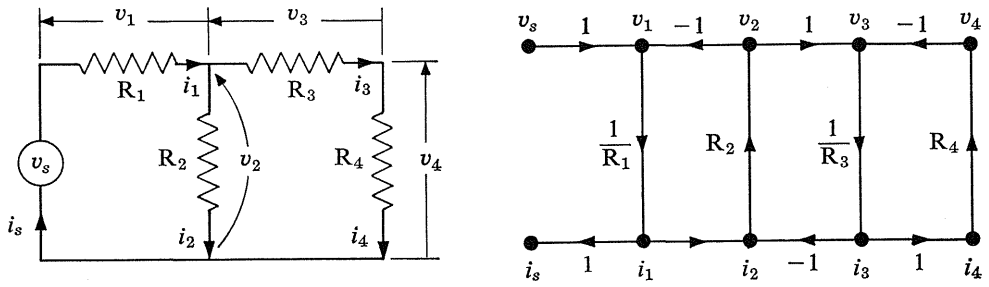


FIGURE 1.33 *Voltage-current relations in a "ladder" circuit.*

flow-graph differs from that previously given in that v_s is now assumed to be known (that is, v_s is shown as a source). The signals at all other nodes (including i_4) therefore depend on v_s. This flow-graph expressed a set of relations between the same signals as before, but the point of view is different. For instance, once v_s is given, the value of v_2 is known, since $v_1 = v_s - v_2$. However, v_2 depends on i_2, which depends on i_1, which depends on v_1. Thus, the simple open cause–effect relationship is here replaced by a circular dependency. The value of a signal depends ultimately on its own value, as is made manifest by the appearance of loops in the graph. (There are three loops in this graph—can you find them?) To solve the equations that are represented by a loop (or feedback) graph such as this, will require the simultaneous solution of the equations.

We next need to consider the general problem of "solving" a flow-graph without writing out the algebraic equations that the graph represents. For instance, suppose that we are interested only in knowing the input current i_s and output voltage v_4, when the input voltage v_s is given. It would be nice to have a method to eliminate all other signals from the graph, so that we could write down the solutions directly by simple inspection. We would like to find A and G_{in} in the reduced graph in Figure 1.34.

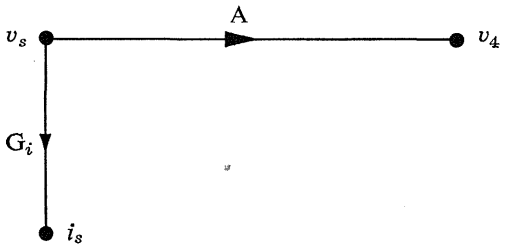
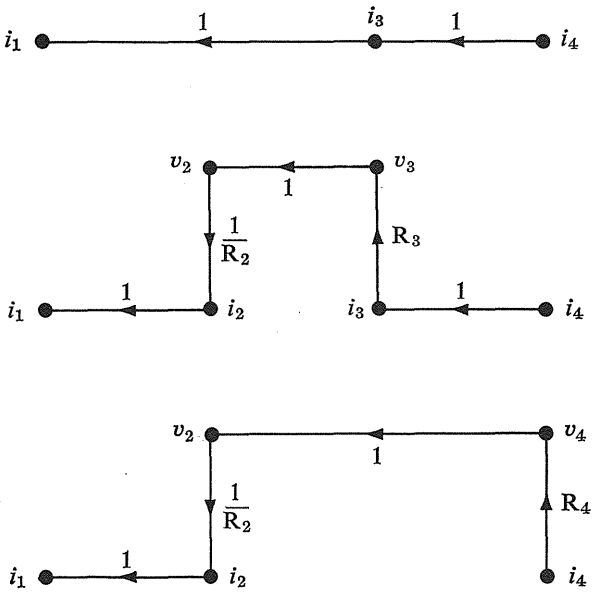


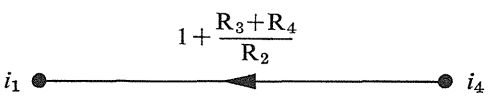
FIGURE 1.34 *A simplified form of the flow-graph of Figure* 1.33.

To see how to do this, we need a set of rules for manipulating the flow-graph directly rather than the more familiar algebraic equations for which it stands. For this development, read now Chapter 2 on Signal Flow-Graphs.

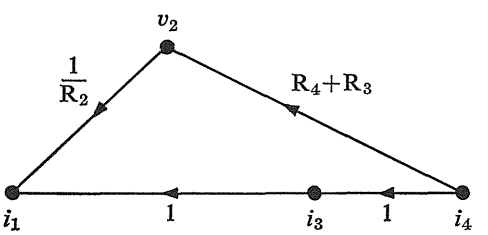
ANSWER TO QUESTION 1.35 There are three open paths from i_4 to i_1, as indicated in the figure. The path operator of the first is 1; of the second, R_3/R_2; and of the third, R_4/R_2. Notice that these three paths are all alternate routes between nodes i_4 and i_1. Even though some portions may be common to two different paths, the paths are different because they are *not identical* throughout.



The signal flow into i_1 is given by the input signal i_4 multiplied by the *sum* of the path transmittances. Thus, insofar as the relation between nodes i_4 and i_1 is concerned, the graph is characterized by the single *graph transmittance* of the operator

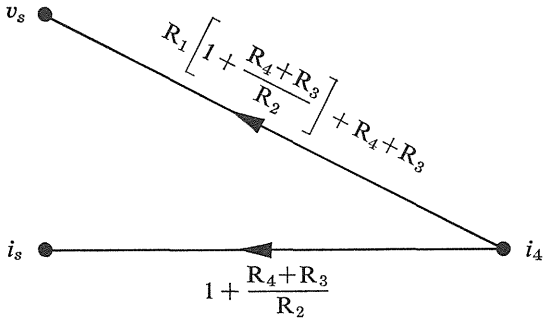


Notice that of the two paths into i_1, one path comes from i_3 and one path comes from v_2.



But the graph operator from i_4 to v_2 has already been found to be $R_4 + R_3$. By combining these two paths we again obtain the same *graph* operator from i_4 to i_1.

ANSWER TO QUESTION 1.36 The transmittances of the single reduced graph are those of the operators shown:



Summary

1. A useful physical observable is characterized by its:

   Invariability: significance is the same throughout physical space and time.

   Generality: applicability to a wide variety of physical systems.

   Additivity: capability of being assigned a numerical value by some measurement process in such a way that numerical addition of these values corresponds to some meaningful physical process of composition.

2. A *signal* is a physical observable to which a unique single number may be assigned at each instant of time by some specified measuring process. (The "meaning" and physical significance of the signal are determined by this measurement process.)

3. There is a correspondence between a physical signal and its numerical value, thus permitting the signal to be described by a function of time. However, a physical signal, such as the temperature, is defined on the *physical* domain, rather than on the domain of real *numbers*. That is, the value of a signal is obtained by performing a physical measurement rather than by evaluating a power series or solving a differential equation. Hence, a signal is a different kind of entity than an ordinary numerical function such as the logarithm, which is defined on the numerical domain. Remember, you can take the logarithm of the temperature, but not the temperature of the logarithm!

4. An *operator* may be used to denote the physical cause–effect relationship between two signals. If the value of the "effect" or output signal at any instant depends only on the value of the "cause" or input signal at the *same instant*, the operator is said to be *static*. If the value of the output at any instant depends upon the value of the input at *earlier instants*, the operator is said to be *dynamic*.

5. Two operators of utmost importance are *scalors* and *delayors*:

   Scalor: The value of the output signal at any instant is a constant scalar multiple of the value of the input signal at the *same* instant.

Delayor: The value of the output signal at any instant t is equal to the value of the input signal at the earlier time $t - T$, where T is the constant time-delay of the delayor.

6. Any operator H is said to be *linear* if it commutes with any scalor A and if the principle of superposition is valid. That is, if for any input signal $f = f_1 + f_2$,

$$f\mathrm{HA} = f\mathrm{AH}$$

and

$$f\mathrm{H} = f_1\mathrm{H} + f_2\mathrm{H},$$

the operator H is said to be *linear.*

7. Any operator H is said to be *stationary* if it commutes with any delayor DELT. That is, if

$$f\,\mathrm{DEL}T\,\mathrm{H} = f\,\mathrm{H}\,\mathrm{DEL}T$$

for any signal f and any delay time T, the operator H is said to be *stationary.*

8. A signal flow-graph is a pictorial representation for a system in which the signals are denoted by *nodes* or *points*, and the operators which represent the contribution of each signal to the others are denoted by *directed branches.* (Detailed and concise definitions for signal flow-graph terms are given in the Summary of Chapter 2.)

9. In a flow-graph for an electrical system, the branches drawn between voltage nodes represent the constraints imposed by Kirchhoff's voltage law; the branches drawn between current nodes represent the constraints imposed by Kirchhoff's current law; and all other branches represent the voltage-current relations for the various circuit elements. These relations may be expressed in many different ways, which are all equivalent to each other. Hence, the flow-graph representation of a system is not unique, although the signals at the various nodes may be identical for every equivalent graph.

NTRODUCTORY YSTEMS AND DESIGN

W. H. Huggins | *Doris R. Entwisle*

THE JOHNS HOPKINS UNIVERSITY

BLAISDELL PUBLISHING COMPANY

A DIVISION OF GINN AND COMPANY

WALTHAM, MASSACHUSETTS • TORONTO • LONDON

2 Signal Flow-Graphs

Traditional mathematical models used to describe engineering structures commonly take the form of sets of algebraic and differential equations. The notation has evolved slowly over the past few centuries. The algebraic part of Descartes' *Geometry*, published in 1637, is still readable; and the notation of the infinitesimal calculus, invented by Leibniz and first published in 1684, has undergone little change within the last century. For many purposes, classical notation may suffice. Those of us who are concerned with the use of mathematical models in the study of engineering systems, however, may wonder whether there is a notation better suited for system representations.

The set of equations that describes the behavior of a system fails to portray the structure of the system-as-a-whole. Each equation is perceptually isolated from the others, and conventional notation does little to connect these pieces together into a coherent whole. Although the entire collection of equations does suffice mathematically to define the structure (in the same sense that a vector can be described by tabulating its components or projections upon some set of coordinates), it nevertheless fails to convey in easily perceived form the structural (or topological) aspects of the system.

A common method of utilizing mathematics in science is to decompose the whole into a sum of its parts. As a limiting example of this method, many laws of the physical world have been expressed in terms of the relations between infinitesimal elements. The solution to any particular physical problem is then effected by assembling these infinitesimal elements (through integration of the appropriate differential equations) so as to satisfy the prescribed boundary conditions. The problems that are readily solved in physics usually pertain to systems that possess very simple structures with homogeneous properties, and the necessity for dealing with complex heterogeneous system structures has not been particularly great in the classical development of mathematical physics. But in man-made systems, homogeneity is the exception rather than the rule. Each discrete element of the system may exhibit a behavior pattern that differs distinctly from that of all other elements, but we are still faced with the necessity for synthesizing these many elements into a composite system having prescribed overall characteristics.

To express this system structure symbolically there has emerged in many fields of modern technology various kinds of flow-graphs ("flow charts," "block diagrams,"

etc.). These diagrams are usually intended to portray not the things that constitute the system, but rather the various operations that the system performs on the stuff that it processes. These flow-graphs are often intended merely to provide a pictorial description of the system to guide the mathematical analysis. Certain ideas and concepts, it turns out, may be communicated much more effectively by use of block diagrams and flow charts than by algebraic forms. Many people find flow-graphs less abstract and more readily identified with the significant properties of the physical systems represented. This suggests that flow-graphs are evolving into a *new kind of notation* that provides a concise, easily visualized description of system structure. Fortunately, flow-graphs can be manipulated and "solved" just as the conventional algebraic and functional symbols, invented several hundred years ago, may be manipulated and solved. The history of mathematics reveals that very often critical breakthroughs follow closely on improvements in notation. Tradition is strong, however. History shows that Arabic numerals were only reluctantly accepted in place of Roman numerals in spite of their conceptual simplicity. And their acceptance came only in response to a strong need. To represent systems with flow-graph notation is a response to a present need, particularly students' needs, since the complexities and interrelationships of the systems are easily apprehended in this form. Flow-graphs form the topic of the present chapter.

Fundamental Definitions

Any system involving simultaneous linear relationships among a number of signals can be expressed in flow-graph form. In this chapter we shall show how relationships among the various signals can be obtained *directly by inspection* of the graph without the necessity for manipulating algebraic equations.* Since these flow-graph techniques will be equally applicable to many linear problems, we shall use a general notation for the signals and operators and their transmittances.

We shall use *lower-case italic* letters to denote signals, such as voltage (e), current (c), force (f), displacement (d), etc. When we are not interested in identifying particular kinds of signals by special letters such as e or c, we shall use the subscripted letter x_i. The subscript can be assigned different values (such as 1, 2, 3, . . . or a, b, c, . . .) to denote different signals of any kind. Previously where signals have not been identified we have already used this notation when x_1 has been the value of the signal at node 1, x_2 has been the value of the signal at node 2, etc. We can also subscript identifying symbols such as e or i, of course. We have earlier talked about e_1, e_2, etc.

We shall use *capital Roman* letters such as A, B, C, . . . , to denote operations on a signal. Thus, the signal x_2 resulting from the operation A on a signal x_1 would be written algebraically as

$$x_1\mathrm{A} = x_2. \tag{2.1}$$

Equation 2.1 states a relation whereby x_2 depends upon x_1 by virtue of the operation A. For instance, x_1 might denote the amount of gasoline in Jones' gas tank, and x_2 the reading of his gas gage. The symbol A would then denote the complex mechanical–electrical–thermal system that converts x_1 into x_2. We can show this relation by denot-

* S. J. Mason and H. J. Zimmerman, *Electronic Circuits, Signals and Systems.* New York: Wiley, 1960, Chap. 4.

ing each signal by a *node* (or heavy dot). The *directed branch* drawn from x_1 to x_2 indicates the asymmetric dependency of the gage reading on the amount of gasoline in the tank. A node such as x_2 with *incoming branches* is called a *dependent* node. A node with *no incoming* branches is called a *source*, or *independent*, node.

We have previously remarked that signals are described by values which depend on the time at which they are measured. The value of the signal x_1 at any instant, t, will be denoted by $x_1(t)$. Likewise, the value of x_2 at this *same* instant will be denoted by $x_2(t)$. Alternatively, we may also express $x_2(t)$ as $x_1\mathrm{A}(t)$ since $x_1\mathrm{A} = x_2$. (The *composite* symbol $x_1\mathrm{A}$ is interpreted as a *single* "*word*" that is merely another name for x_2. It does not necessarily mean x_1 multiplied by A. It means "the result obtained when x_1 has been operated upon by A.")

However, to simplify the initial discussion, we shall use *scalor* operators. A scalor, such as A in Figure 2.1, amplifies the signal x_1 by some constant factor A. The equation

FIGURE 2.1 *A causal relation between two signals.*

$x_1\mathrm{A} = x_2$ implies that $x_2(t) = Ax_1(t)$ for any t. Because the algebraic form of a scalor operator is so similar to its transmittance expression, in this chapter, we shall often use the word "transmittance" to refer to the operator expression as well as to the numerical value of its transmittance.

Consider next two signals x_2 and x_3, each of which is dependent on x_1 in accord with the expressions

$$x_1\mathrm{A} = x_2, \tag{2.2a}$$

$$x_1\mathrm{B} = x_3. \tag{2.2b}$$

To use the same example, x_1 might denote the gas added to the tank, x_2 the indication of the gas gage, and x_3 the increase in the weight of the automobile. We may construct a signal flow-graph describing this situation by including a third node to represent x_3, such as Figure 2.2. This graph is merely another representation for Equations 2.2a

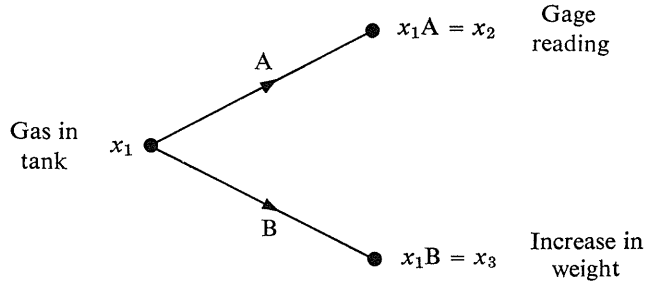


FIGURE 2.2 *The signal at a node is operated on independently by each outgoing branch.*

and b. In those equations, it was necessary to write x_1 twice (once in each equation), but in the graph we show the node only once. More generally, the signal x_1 acts through each *outgoing* branch, and the fact that two (or more) branches originate from a given node in no way alters the signal at that node. Hence, in Figure 2.2, the inclusion of the lower branch, B, for x_3 does not affect the signal x_2.

The signal at a dependent node is composed of the total signal flow into that node. Thus, a signal x_4 might depend upon x_2 and x_3 in accord with the relations

$$x_2C + x_3D = x_4. \tag{2.3}$$

We can write Equation 2.3 because of the *additive* property that we assume is possessed by all of our observables.* We may construct a signal flow-graph to represent x_4 as shown in Figure 2.3.

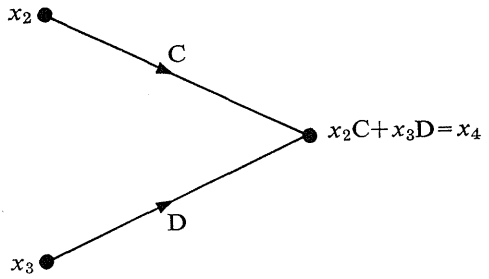


FIGURE 2.3 *The signal at a dependent node is composed of the total signal flow into that node.*

To summarize the three basic definitions (1, 2, 3,) and the three rules (A, B, C) governing signal flow-graphs:

1. The signals (i.e., observables) are denoted by nodes.
2. The additive contribution of one signal to another is denoted by a *directed branch* from the "causal" node to the "dependent" node.
3. A node with no incoming branches is called a *source node.* It represents a signal which must be independently specified. A node with at least one incoming branch is called a *dependent node* because it represents a signal that depends upon other signals in the graph.

A. The signal at any *dependent* node is composed of the total signal flow *into* that node from all *incoming* branches. (The signal at each *source* node is assumed to be specified independently.)
B. The signal at a node is operated upon independently by *each outgoing* branch. (That is, *outgoing* branches in no way affect the signal at a node.)

* We are restricting our attention to a particular class of mathematical models. It *happens* that simple additivity characterizes many of the relations that exist in the real world. Notice that this is not as restrictive as might be imagined at first. For instance, if node signals are measured on a logarithmic scale, the additive property would be a model for a multiplicative situation.

C. The signal at the input end of a branch is transformed or modified in some specified way as it passes through the branch. (The branch therefore represents a prescribed operation upon the input signal to obtain the transformed signal at its output. Thus, in general, the branch denotes an *operator*.)

Make sure that you know each of these statements well enough so you can repeat it after you turn the page. These statements are illustrated symbolically by Figure 2.4, which shows the signal "flow."

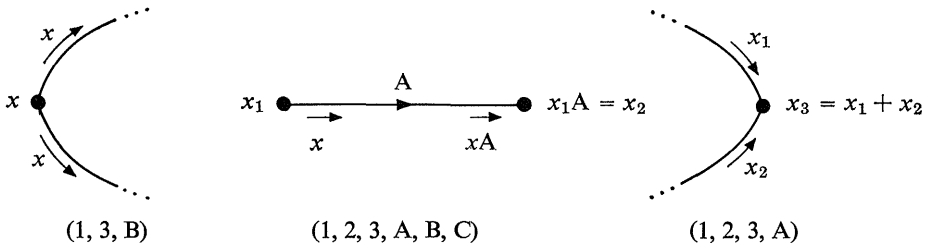


FIGURE 2.4 *The numbers and letters below each diagram refer to the particular statements required to write and interpret each diagram.*

Elementary Transformations

A number of elementary transformations follow directly from these statements which permit reduction and simplification of a graph. These are illustrated by Figure 2.5. The validity of these transformations may be proved easily by showing that a given source signal will produce the same signal values at the sink nodes in either graph. Note that the signal at a node is composed of all incoming signals. Hence, the value of the node signal *at any instant* is equal to the sum of the values of all incoming signals *at that same instant.*

QUESTION 2.1 Written below are several sets of algebraic equations, each of which is equivalent to one of the signal flow-graphs shown in Figure 2.5. Which graph corresponds to each set of equations? (First draw the graph for each set of equations and *then* compare with the diagrams given previously.)

(1) $x_4CA = x_1$
$x_4CB = x_2$

(2) $x_1A + x_2B = x_3$
$x_3C = x_4$

(3) $x_1A + x_1B = x_2$

(4) $x_3A = x_1$
$x_3B = x_2$
$x_4C = x_3$

(5) $x_1A = x_3$
$x_3B = x_2$

QUESTION 2.2 What is a set of algebraic equations which express the same signal relations portrayed by the following flow-graph? (In this and some of the following graphs, we may label the node x_j simply by its index j). (Answer)

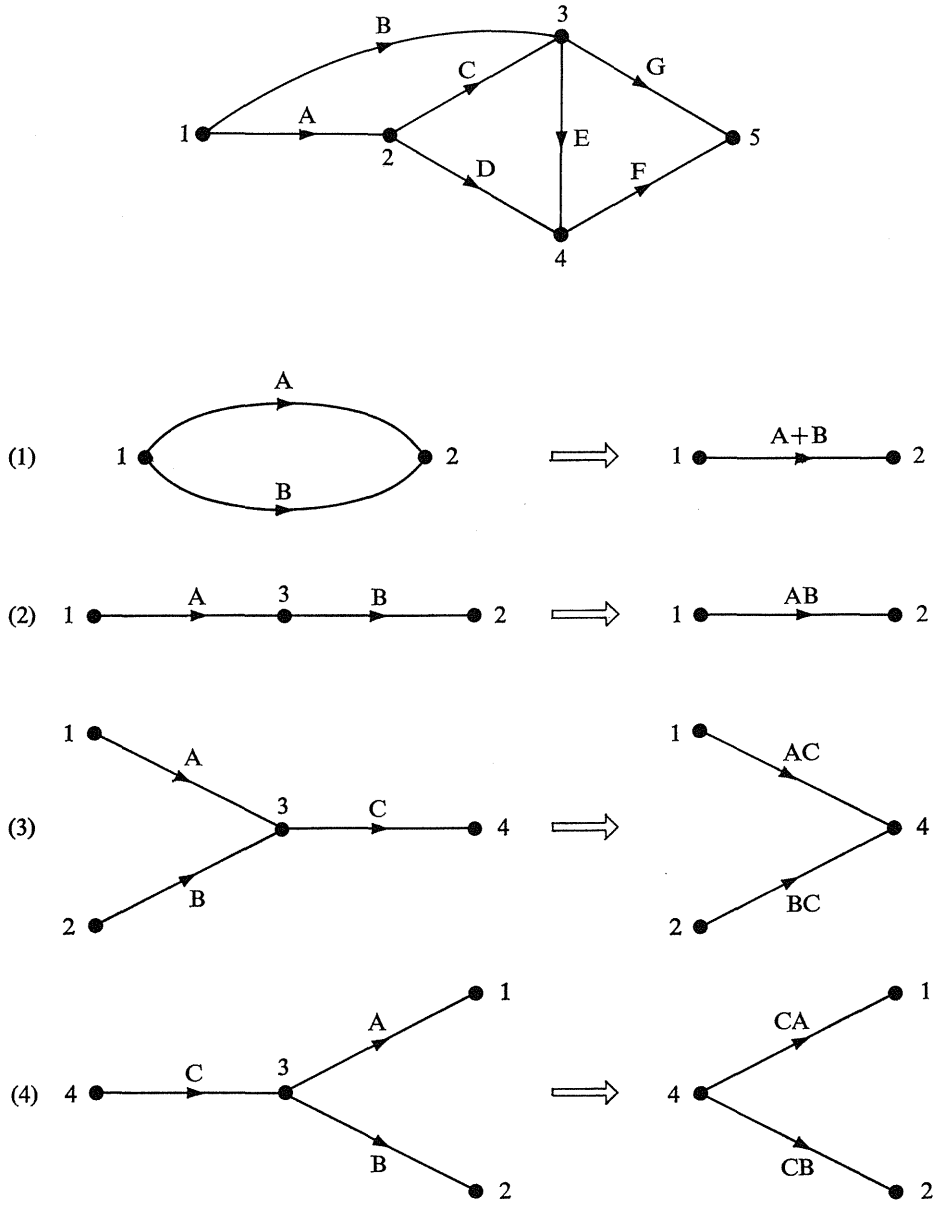


FIGURE 2.5 *Elementary reductions.*

QUESTION 2.3A In writing simultaneous equations it is sometimes useful to use "filing-cabinet numbers" called *matrices*. Is there any way in which the following arrangement of symbols may be interpreted so as to be equivalent to the algebraic equations that you produced in Question 2.2? (Answer)

	x_1	x_2	x_3	x_4	x_5			
1	1						1	x_1
2	A						2	x_2
3	B	C				=	3	x_3
4		D	E				4	x_4
5			G	F			5	x_5

QUESTION 2.3B What signal flow-graph corresponds to the matrix equations shown below? (Answer)

x_1	x_2	x_3	x_4		
			1		x_1
	1				x_2
A	B			=	x_3
		C			x_4

With reference to graph 3 of Figure 2.5, consider next the effect of making the signal at node 1 equal to the signal at node 4. This may be accomplished in flow-graph symbolism by drawing from node 4 to node 1 a branch having unity transmittance; node 1 thus ceases to be a *source* and becomes a *dependent* node by virtue of its having input branches as shown in Figure 2.6.

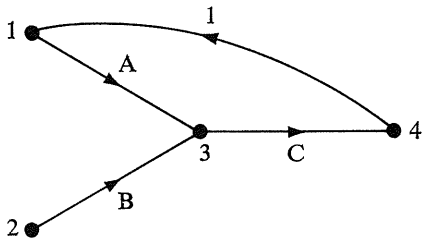


FIGURE 2.6

By using the elementary transformations described above, it is now possible to *absorb* node 3 and *contract* the "single-loop" graph to a simpler form. Two graphs are *equivalent* as long as signals are the same at corresponding nodes in the two graphs. Nodes may be eliminated, however, without changing the signals at the remaining nodes. It is meaningful therefore to speak of a *reduced graph* (in which one or more of the original nodes have been eliminated) as equivalent to the original graph if the signals at the remaining nodes are the same as in the original graph. For instance, by applying the elementary reduction 3 of Figure 2.5 to the graph shown in Figure 2.6, we find

the results depicted in Figure 2.7. In the graph at the left in Figure 2.7, node 3 has been eliminated to obtain the reduced graph shown at the right. In the left graph, the corresponding algebraic equations are

$$x_1A + x_2B = x_3,$$
$$x_3C = x_4,$$

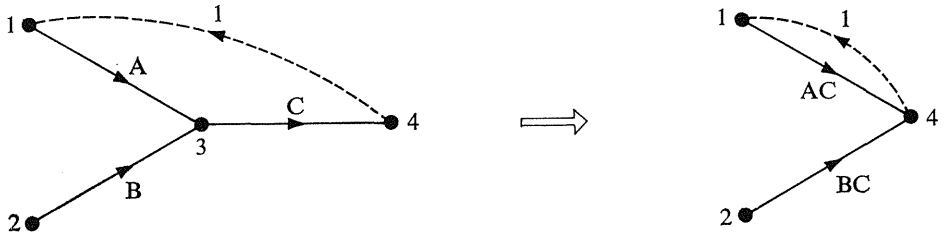


FIGURE 2.7 *Absorption of node* 3.

whence, by replacing x_3 by its equivalent expression given by the first equation,

$$x_1AC + x_2BC = x_4.$$

But this equation corresponds to the right graph. Hence, at the remaining nodes 1, 2, and 4, the two graphs are equivalent.

ANSWER TO QUESTION 2.2

x_1 (source)

$x_1A = x_2$

$x_1B + x_2C = x_3$

$x_2D + x_3E = x_4$

$x_3G + x_4F = x_5$

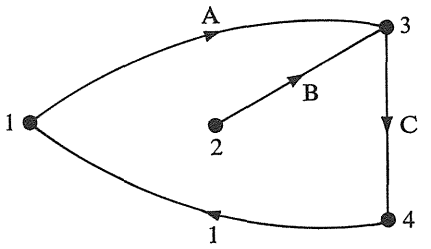
ANSWERS TO QUESTION 2.3

A) 1. Each *row* of the matrix contains the operators that act on the signals shown at the top of each *column.*

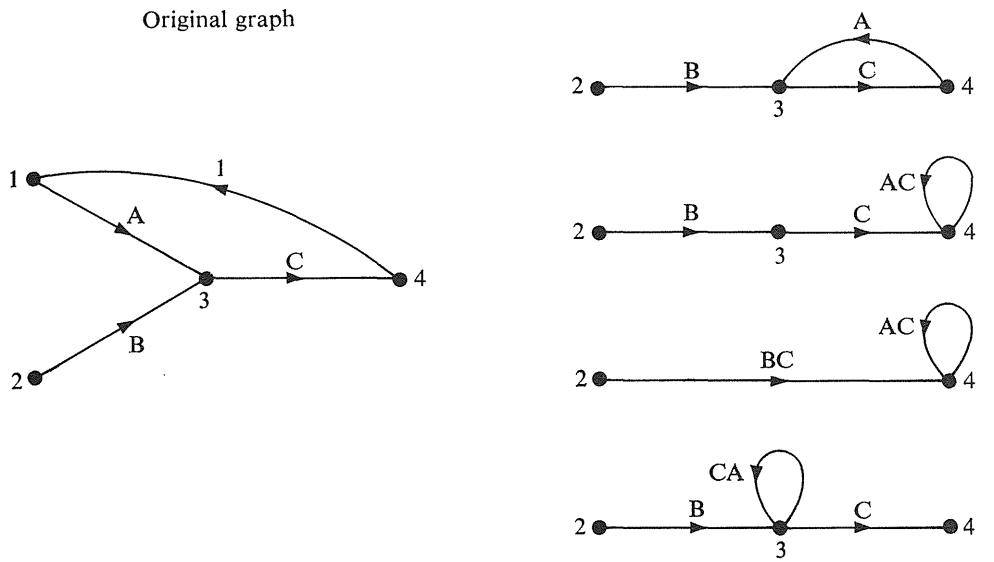
2. Each row thus represents one equation (compare with the answers given to Question 2.2 above).

3. Note that *outgoing* branches comprise the *columns*, and *incoming* branches comprise the *rows.*

B)



QUESTION 2.4 The figure shows several reductions of the signal flow-graph just considered. All except one of these reduced graphs are "equivalent" to the original graph in the sense just described. Which graph is *not* equivalent? (Answer)



The discussion of Question 2.4 illustrates a general procedure for *proving* whether or not two signal flow-graphs are equivalent at the nodes which they both have: Simply express the relations between the node signals in algebraic form. If identical expressions can be obtained for the signals in both graphs, then the graphs are equivalent. Another possibility for testing two graphs for equivalence is to see whether they both accomplish the *same sequence of linear operations*. For instance, any continuous path through the graph corresponds to a possible sequence of linear operations upon a signal "flowing" over that path. Thus, the left- and right-hand graphs shown in Figure 2.8 produce the

FIGURE 2.8 *Equivalent graphs.*

same sequence of operations on a signal as it flows *over all possible paths* from 2 to 4. Since the loop may be traversed 0, 1, 2, . . . times, there are, in effect, an infinite number of paths through either of these graphs. For instance, the left-hand graph has paths corresponding to the following sequence of operations

$$x_2\mathrm{B}[1 + \mathrm{CA} + \mathrm{CACA} + \mathrm{CACACA} + \cdots]\mathrm{C} = x_4,$$

whereas the right-hand graph has paths corresponding to the operation

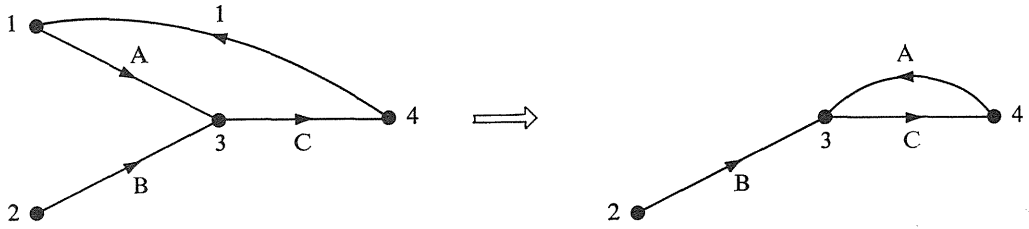
$$x_2\mathrm{BC}[1 + \mathrm{AC} + \mathrm{ACAC} + \mathrm{ACACAC} + \cdots] = x_4.$$

By multiplying these sequences out, we see that *both* are of the form

$$x_2[\mathrm{BC} + \mathrm{BCAC} + \mathrm{BCACAC} + \cdots] = x_4.$$

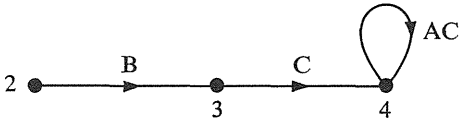
Hence, both graphs represent the same sequence of operations on x_2 to yield x_4. They are therefore equivalent.

ANSWER TO QUESTION 2.4 In reducing a graph, you should recall that the signal at a node is determined by the incoming branches. Outgoing branches may be altered without changing in any way the signal at that node (except that changing an outgoing branch may produce in a roundabout way a change in the signal entering via the incoming branches when feedback is present). The reduced graphs are equivalent to the original graph *only* if the signals at corresponding pairs of retained nodes are equal. The first reduction follows immediately from $1 \cdot \mathrm{A} = \mathrm{A}$:



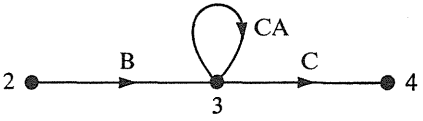
First reduction

Hence, the only reduced graphs about which there could possibly be any doubt are the second and fourth.



Second reduction

and



Fourth reduction

The graphs will be correctly transformed provided the relationships that they depict are consistent with those of the first reduced graph. Let us examine the value of the signal at node 3 in each of the three graphs shown above.

$$\text{First graph:} \quad x_3 = x_2\mathrm{B} + x_4\mathrm{A},$$

$$\text{Second graph:} \quad x_3 = x_2\mathrm{B},$$

$$\text{Fourth graph:} \quad x_3 = x_2\mathrm{B} + x_3\mathrm{AC}.$$

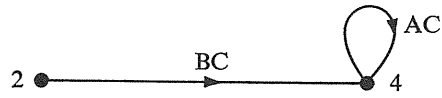
At first glance, it would appear that each graph gives a different value for x_3. But it will be seen that for the first graph, $x_4 = x_3\mathrm{C}$, whereupon it follows that the values for x_3 given by the first and fourth graphs are actually equivalent. However, *neither* of these values for x_3 is equivalent to the x_3 of the second graph, *which is therefore in error.* Let

it be noted, however, that the third reduced graph of the figure in Question 2.4 is correct, since it avoids making any implications about x_3, and merely requires that

$$x_4 = x_2\mathrm{BC} + x_4\mathrm{AC},$$

which is seen to be equivalent to the relationship

$$x_4 = x_3\mathrm{C} = [x_2\mathrm{B} + x_4\mathrm{A}]\mathrm{C}.$$



To illustrate this process with a slightly more involved example, let us eliminate nodes 2, 3, and 4 from the graph shown in Figure 2.9. The signals at each node are:

at node 1: x_1 (source),
at node 2: $x_2 = x_1\mathrm{A} + x_3\mathrm{B} + x_4\mathrm{C}$,
at node 3: $x_3 = x_5\mathrm{E}$,
at node 4: $x_4 = x_5\mathrm{F}$,
at node 5: $x_5 = x_2\mathrm{D}$.

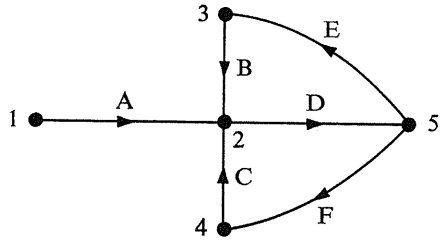


FIGURE 2.9

To eliminate nodes 3 and 4, we need only remember that the transmittance of two branches in cascade is the product of the transmittance of each of the branches, as in Figure 2.10. The signals present now are:

at node 1: x_1 (source),
at node 2: $x_2 = x_1\mathrm{A} + x_5\mathrm{EB} + x_5\mathrm{FC}$,
at node 5: $x_5 = x_2\mathrm{D}$.

This reduced graph is equivalent to the original graph.

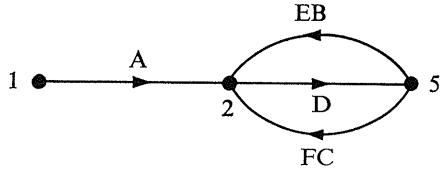


FIGURE 2.10

The two branches which return a signal from x_5 to x_2 have transmittances that add, as is clear from the second equation above. We can just as well make the graph a little simpler to look at, as Figure 2.11. The signal at node 2 is still

$$x_2 = x_1A + x_5(EB + FC).$$

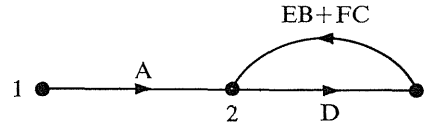


FIGURE 2.11

To obtain an equivalent graph we eliminate node 2, using elementary reduction 3 of Figure 2.5, as shown in Figure 2.12. It does not matter in which "direction" we write

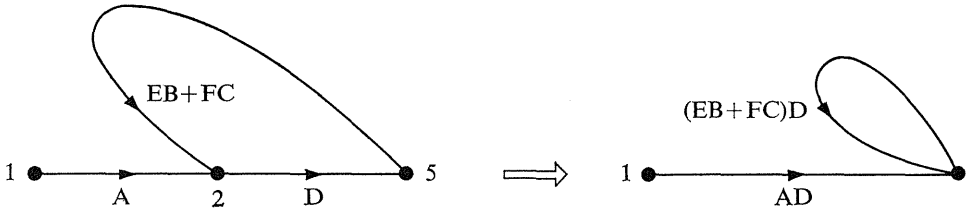


FIGURE 2.12

the arrow on the self-loop at node 5 of the right-hand graph. The signal at node 5 is

$$x_5 = x_1AD + x_5(EB + FC)D.$$

We may show the equivalence by returning to the original graph and expressing signal x_5 in terms of x_1 and x_5 itself.

$$x_2 = x_1A + x_3B + x_4C,$$

$$x_3 = x_5E,\ x_4 = x_5F,\ x_5 = x_2D,$$

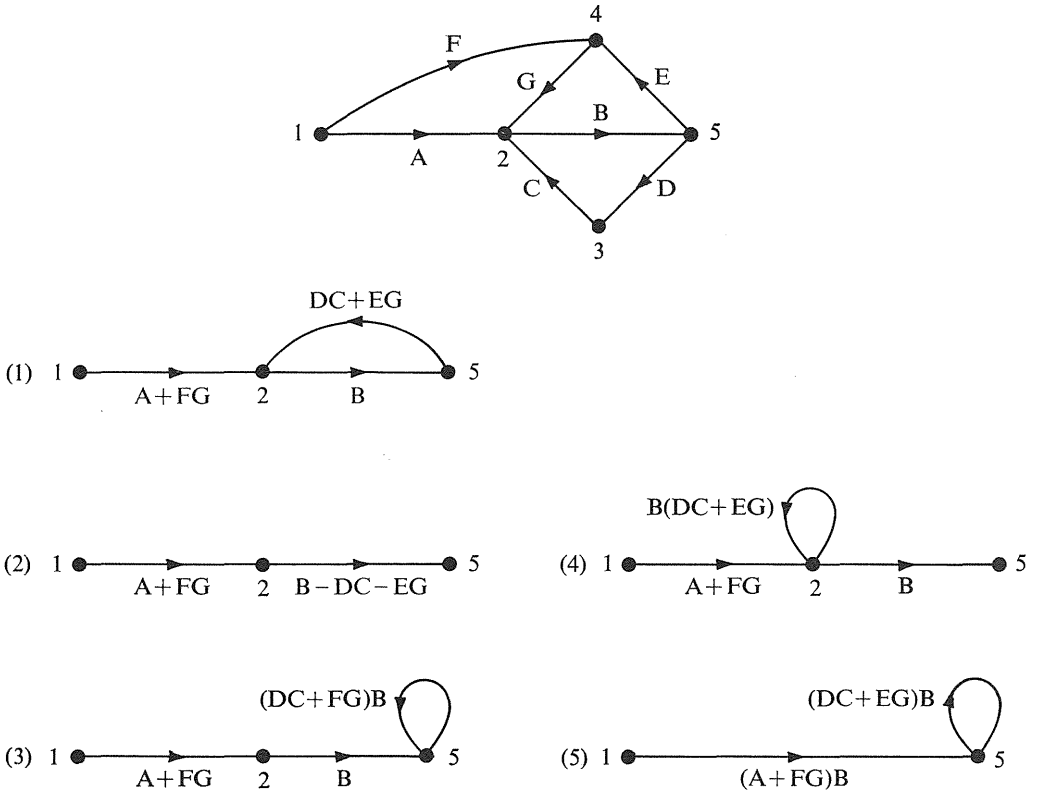
$$x_5 = (x_1A + x_3B + x_4C)D,$$

$$x_5 = x_1AD + x_3BD + x_4CD,$$

$$x_5 = x_1AD + x_5(EBD + FCD).$$

This is identical to the x_5 obtained from the reduced graph.

QUESTION 2.5 Check the reduced graphs, 1 to 5, of the accompanying graph, and determine which are correct reductions to the original graph. (Answer)



Effect of Self-Loops

The reductions you have just studied show that a self-loop may emerge when a graph is reduced. A clear understanding of the effect of a self-loop at a node will be of great importance in all that follows.

Consider the graph shown in Figure 2.13. The signal at node 3 is the *sum of the signals entering the node.* As illustrated by the diagram, x_1 is clearly entering the node,

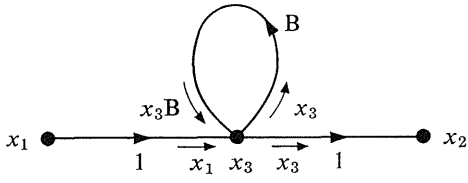


FIGURE 2.13 *A simple self-loop.*

and because of the self-loop on node 3, you see that x_3B is entering the node. Since x_3 is the node signal and x_3B is furnished by the self-loop, only the difference $x_3(1 - B)$

must be supplied from the source. For this reason the quantity (1 − B) is called the *return difference* in feedback theory.

$$x_1 + x_3B = x_3.$$

Hence,

$$x_3(1 - B) = x_1.$$

In particular, when B = 1, the return difference becomes zero, which is to say that *all* of the input signal needed to create a signal x_3 at the node is "fed back" as a consequence of the signal itself, and *no* external *signal* at all is required to sustain the signal x_3. This peculiar condition governs the natural behavior of a system, and plays a vital role in the study of stability and other aspects of the dynamic behavior of linear physical systems.

Now that you have grasped the significance of the effect of a self-loop, you should be able to determine the equivalent transmittance of a path that contains a node having a self-loop, such as that shown in Figure 2.14.

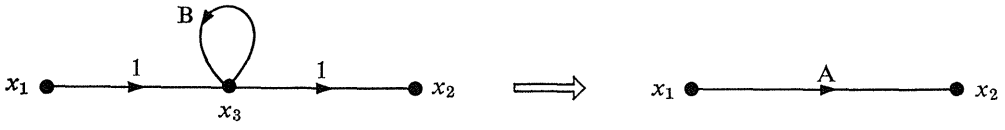


FIGURE 2.14

The equivalent transmittance A is obtained by first noting that:

$$x_1 + x_3B = x_3,$$

$$x_3 = x_2.$$

Rearrangement of the first equation yields

$$x_3(1 - B) = x_1 \quad \text{or} \quad x_1(1 - B)^{-1} = x_3,$$

whence

$$x_1(1 - B)^{-1} = x_2 \quad \Rightarrow \quad x_1A = x_2.$$

Therefore,

$$\underline{(1 - B)^{-1} = A} \qquad \textit{IMPORTANT RESULT!}$$

ANSWER TO QUESTION 2.5 The correct reductions are Nos. 1, 4, and 5.

Graph 2 is wrong because it ignores the effect of node 5 on node 2. Furthermore, the transmittance shown in the graph from 2 to 5 is incorrect.

Graph 3 is not equivalent to the original graph because the signal at node 2 is not the same in the two graphs, even though the signals at nodes 1 and 5 are the same. If in this graph, node 2 were *relabeled* so as to denote a *different* signal than x_2, then since the signals at the corresponding x_1 and x_5 are the same, the two graphs would be equivalent at these nodes.

The effect of a self-loop, B, *at any node is to modify an incoming signal by the operator* $(1 - \mathrm{B})^{-1}$ *as the signal passes into the node.* This result is valid for *all values* of B, since your derivation of this result was independent of the magnitude of B.

Another interpretation of this result considers the many different routes by which the signal can find its way from the source x_1 to the dependent node x_2 in Figure 2.15.

$$x_1[1 + \mathrm{B} + \mathrm{B}^2 + \mathrm{B}^3 + \cdots] = x_2,$$

$$x_1\left[\frac{1}{1 - \mathrm{B}}\right] = x_1(1 - \mathrm{B})^{-1} = x_2.$$

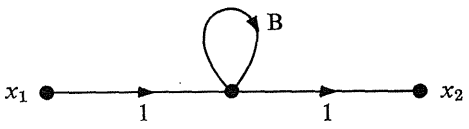


FIGURE 2.15

Although there is only *one open* path from x_1 to x_2, there are an infinite number of *paths* (corresponding to 0, 1, 2, 3, . . . circulations around the self-loop). These successive paths have, respectively, the transmittances 1, B, B^2, B^3, ... Since the paths are all originating from x_1 and terminating on x_2, they are all effectively in parallel and the sum of the transmittances yields the infinite geometric series which sums to $1/(1 - B)$, thus agreeing with our previous result.

When several branches enter and leave the node, it is helpful to remember that the self-loop serves to amplify each *incoming* signal by the operator $(1 - \mathrm{B})^{-1}$. This effect may be shown explicitly by separating the incoming from outgoing branches and introducing a new branch having a transmittance of $(1 - B)^{-1}$ according to the scheme in Figure 2.16. This transformation is illustrated by the reduction in Figure 2.17. Ob-

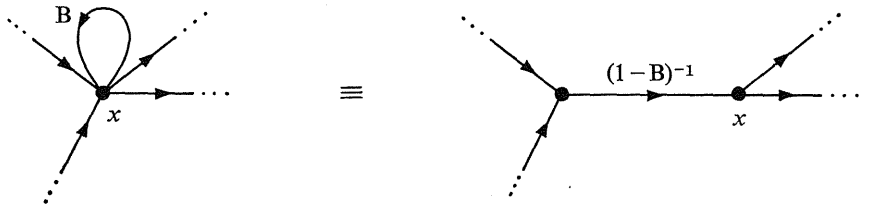


FIGURE 2.16 *A transformation for eliminating a self-loop.*

serve that the *output* node of the branch $(1 - \mathrm{B})^{-1}$ corresponds to the node 5 having the self-loop in the original graph.

Using the rule we have just learned, let us show that a graph we have considered before is equivalent to its two reductions shown at the right in Figure 2.18. Express the signals at each node in terms of the signals at the source.

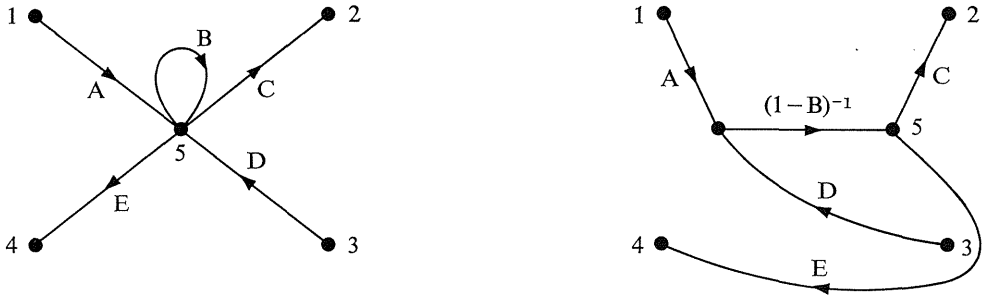
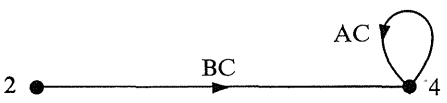


FIGURE 2.17

Graph 1 Graph 2



Graph 3

FIGURE 2.18

For graph 1

$$x_3C = x_4,$$

$$x_2B + x_4A = x_3 = x_2B + x_3CA,$$

$$\therefore\ x_2B(1 - CA)^{-1} = x_3.$$

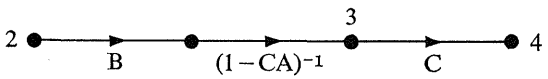
For graph 2

$$x_2B + x_3CA = x_3,$$

$$\therefore\ x_2B(1 - CA)^{-1} = x_3,$$

which agrees with graph 1.

Also, by the transformation of Figure 2.16 which also agrees with graph 1.



or

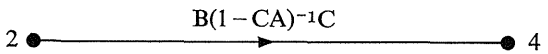


FIGURE 2.19

Similarly, for graph 3.

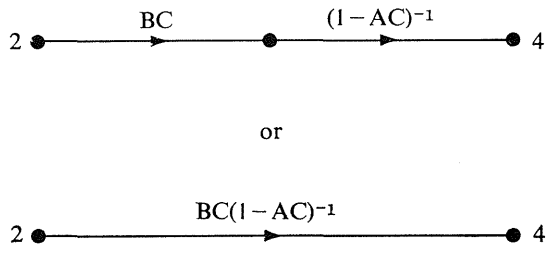


FIGURE 2.20

It is not difficult to see that this signal agrees with the signal at x_4 in the original graph if the operators commute, that is, if AC = CA, and $C(1 - AC)^{-1} = (1 - CA)^{-1}C$. In general, operators do not commute, and to show graphs 2 and 3 equivalent for this most general case, we require that $(1 - CA)^{-1}C = C(1 - AC)^{-1}$. The only assumption that need be made is that the *inverse* of C exists. By this we mean an operator C^{-1} which when combined with C in either order yields the identity operator:

$$CC^{-1} = C^{-1}C = 1.$$

Also, a simple relationship involving the inverse of an operator which is expressed as a product of two invertible operators will be useful. Given DE = F, we wish to show $E^{-1}D^{-1} = F^{-1}$. Here,

$$FF^{-1} = DEE^{-1}D^{-1} = DD^{-1} = 1,$$
$$F^{-1}F = E^{-1}D^{-1}DE = E^{-1}E = 1.$$

Now, to show $(1 - CA)^{-1}C = C(1 - AC)^{-1}$, we rewrite each expression as follows:

$$(1 - CA)^{-1}C = [C(C^{-1} - A)]^{-1}C = (C^{-1} - A)^{-1}C^{-1}C = (C^{-1} - A)^{-1},$$
$$C(1 - AC)^{-1} = C[(C^{-1} - A)C]^{-1} = CC^{-1}(C^{-1} - A)^{-1} = (C^{-1} - A)^{-1},$$

which are identical expressions.

Node Absorption and Graph Reduction

The process of eliminating nodes from a graph corresponds to the elimination of unknown variables from the set of equations which the graph represents. This reduction is useful in analyzing models of physical systems because the reduced graphs obtained along the way express relations between the retained nodes which exhibit the same signals as in the original graph. Thus, we are able to replace a *complicated* system, involving perhaps many signals in which we are not really interested, with a simpler system exhibiting a few relations among the signals of primary interest.

For instance, to construct a detailed model of a hi-fi amplifier, such as that in your phonograph, you need to consider initially dozens of different voltages and currents associated with the electrical resistors, capacitors, and transistors inside the box. Yet,

you may ultimately be interested only in the *input–output characteristics*, which relate the voltage and current at the input terminals to the output of the amplifier. By first constructing a graph of the system, you may eliminate signals in which you are not interested, and obtain equivalent relations between the remaining signals. The reduction process, if carried to the limit, leads to a graph having only *source* and *sink* nodes. In such a graph, all intervening relationships have been eliminated and replaced by an equivalent set of *direct relations* between the independent and dependent signals (similar to that of Figure 2.1).

By using the transformations shown in Figures 2.5 and 2.16, you can now absorb any node in a graph whose branches denote *linear* operations. The most useful of these transformations are the following three:

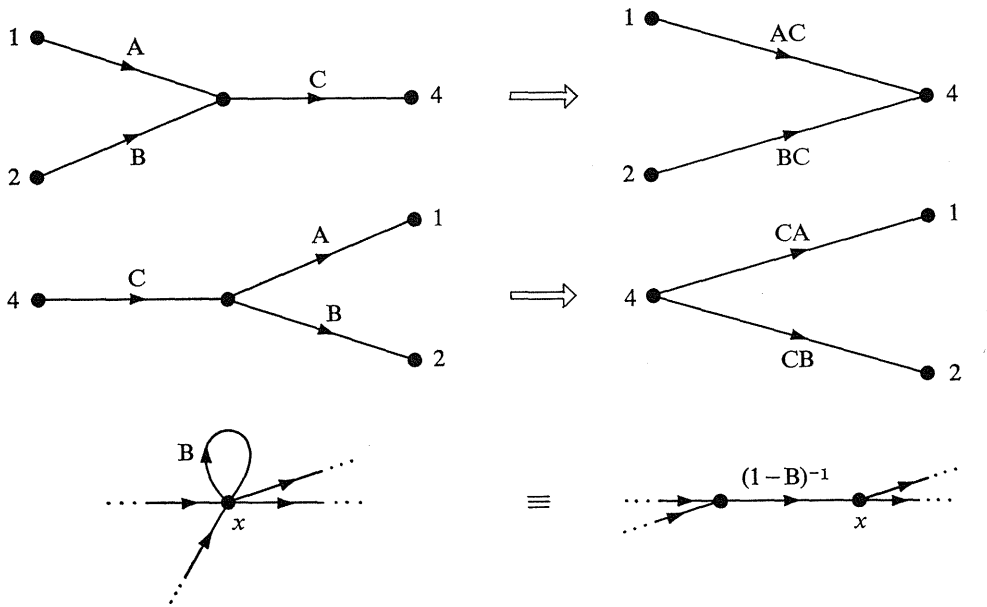


FIGURE 2.21 *Useful transformations for graph reduction.*

Let us apply these transformations to eliminate node x_3 and thus reduce the graph at the left below to its simplest form involving the single equivalent operator between x_1 and x_2, shown at the right in Figure 2.22.

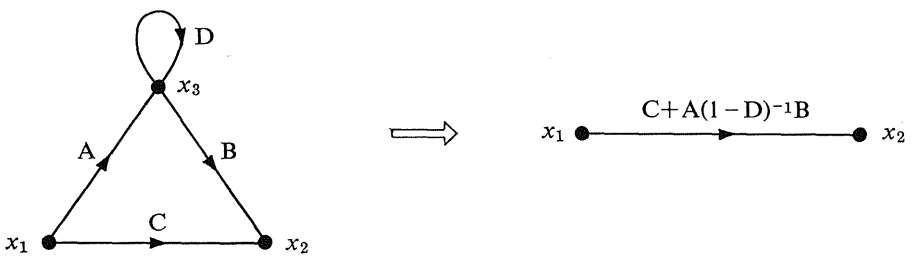


FIGURE 2.22

First, by Figure 2.16 we eliminate the self-loop at x_3, as in Figure 2.23. Since the two paths are in parallel, they may be combined (Figure 2.5(1)) to obtain the reduced graph shown in Figure 2.23.

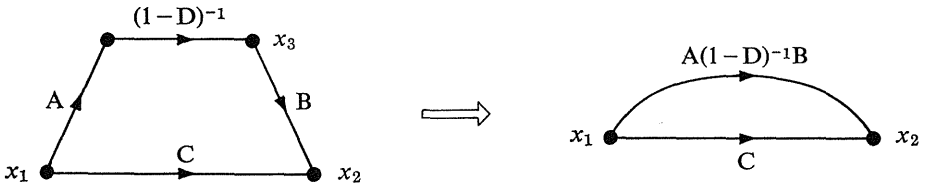


FIGURE 2.23

As a second example, we may consider a somewhat simplified flow-graph representation of a hi-fi amplifier. It consists of three amplifier stages, A, B, and C, connected in cascade so that the output signal of one serves as the input signal of the next. Furthermore, feedback has been incorporated over two of the stages, as well as over the whole amplifier, so as to reduce "hum" and distortion. The flow-graph model of this amplifier appears as Figure 2.24.

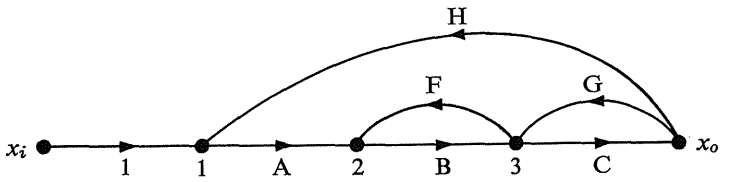


FIGURE 2.24 *A flow-graph of an amplifier.*

Let us eliminate node 3, by use of the transformation shown in Figure 2.25. (The graph of Figure 2.25 gives a generalization of the transformation of Figures 2.5(1) and (2). It is unrelated to Figure 2.24.) You will recall from the discussion in Chapter 1 that signals tend to "find their way" through the graph over all permissible paths. Thus, considering only the signals "resulting" from x_1, we see that x_1 produces $x_3 = x_1B$, which in turn produces $x_4 = x_3C = x_1BC$ and $x_5 = x_3F = x_1BF$. Thus, the contribution to x_4 due to x_1 is represented by the branch BC in the right-hand graph in

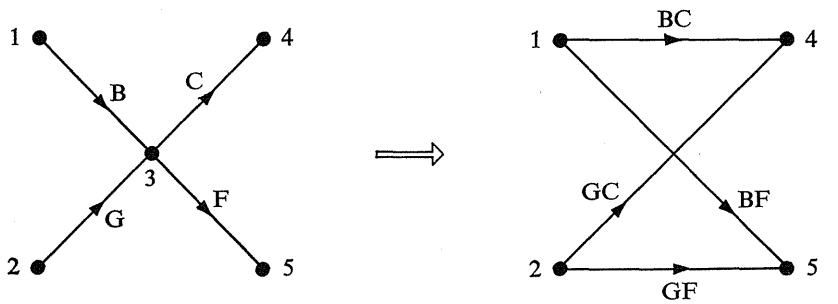


FIGURE 2.25 *A more general transformation.*

Figure 2.25. These branches are equivalent to the two cascaded branches needed to go from x_1 to x_4 (or x_5) in the left-hand graph. Before proceeding, you are urged to make certain that you thoroughly understand why this transformation is valid. (It may help to write out fully the algebraic equations corresponding to the graph.) When we apply this transformation to the graph of Figure 2.24, a new graph, Figure 2.26, is obtained in which node 3 no longer appears.

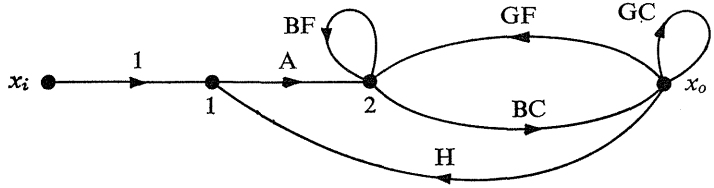


FIGURE 2.26 *Node 3 is eliminated.*

Now we have *two* self-loops. But these can be eliminated by using the transformation of Figure 2.16. Observe carefully that a self-loop affects only the *incoming* branches of a node, and that its effect may be accounted for by temporarily introducing a new node, to receive all incoming branches, which is connected to the original node and its outgoing branches by an operator of the form $(1 - B)^{-1}$, as illustrated in Figure 2.16.

QUESTION 2.6 Draw a transformed graph having the same nodes and node signals as the graph of Figure 2.26, but no self-loops. (Answer)

QUESTION 2.7 Next, eliminate node 2. (Answer)

QUESTION 2.8 Reduce the graph found in the answer to Question 2.6 so as to express in terms of M, N, and H the equivalent operator K relating input and output. (Answer)

Commutative Operators

In most of the foregoing development we have been careful to preserve the order in which successive operations are applied to a signal. For instance, if A and B denote two operations that vary with time, then the combined operation AB is *not* generally the same as BA. In describing time-varying systems, we must take care to preserve the *correct order* of the operations.

However, in this book we shall be concerned primarily with operations which do not change with time. We shall speak of such operators as *time invariant* or *stationary*. This definition may be a little confusing at first. It does *not* mean that we will be ignoring time-varying signals. A stationary operator will produce the same effect on a time-varying signal regardless of when the signal occurs, whereas a nonstationary (time-varying) operator will have a different effect at different times. These notions will be discussed more fully in a later section.

When the operators A and B are *linear* and *stationary*, the combined operation AB is the same as BA and we may simplify the writing of our equations considerably by writing an expression such as $(1 - B)^{-1}$ as an ordinary fraction, $1/(1 - B)$. Thus, if A, B, and C are *linear stationary* operators, the *overall* relation *between* x_i and x_0 will be equivalent in each of the graphs in Figure 2.27.

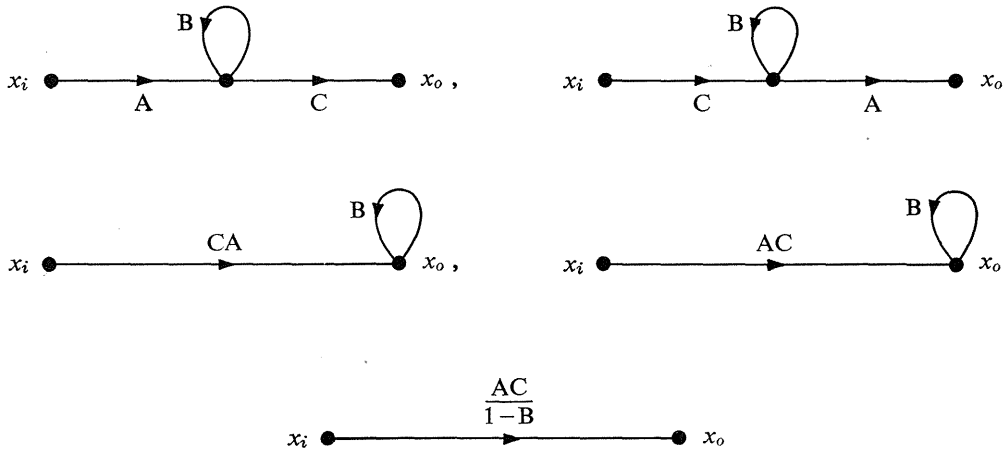


FIGURE 2.27

The fact that the order of the operators may be reversed without altering the overall result is of great practical engineering significance. An example is illustrated by the design of your television receiver. Here, it was found that to ensure the sharpest possible video image, another operation should ideally be performed upon the video signal, after it had been amplified in the television receiver, to compensate for the "phase distortion" which accompanies the amplification operation. Unfortunately, the circuitry to perform this phase-equalizing operation would increase the cost of *each* receiver. The fact that most of the various signal-processing operations in a television link from studio to television receiver in your room are *linear* and *stationary* allows us to use a *single* phase equalizer *at the transmitter* to accomplish the *same* overall result as if we had installed an

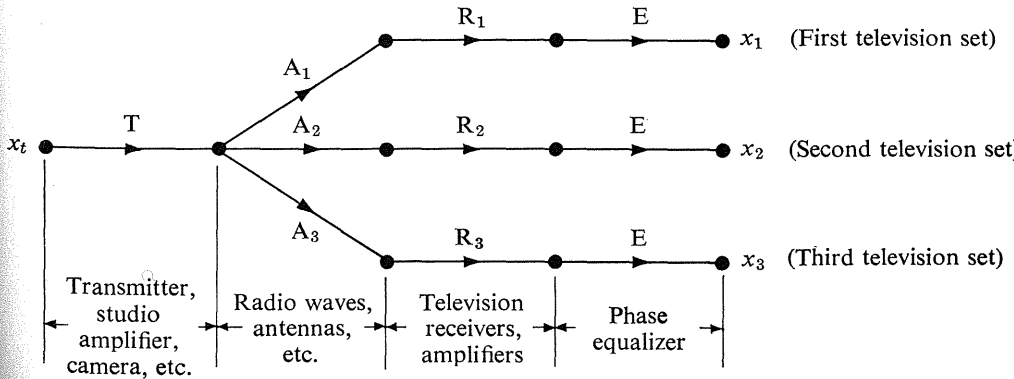


FIGURE 2.28 *A phase equalizer is installed in each television set.*

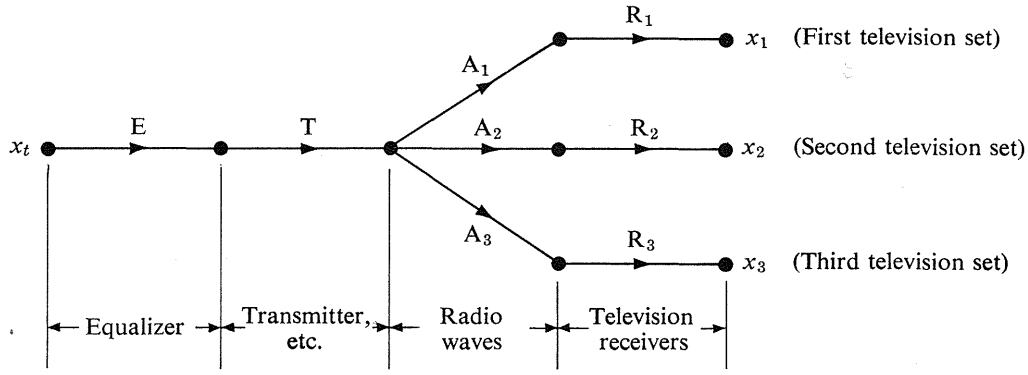
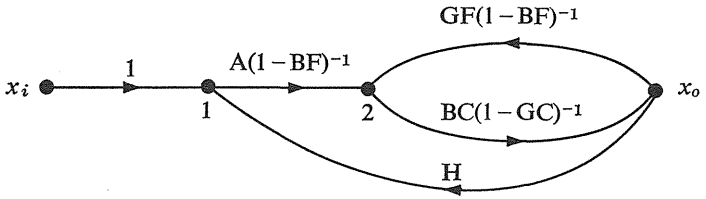
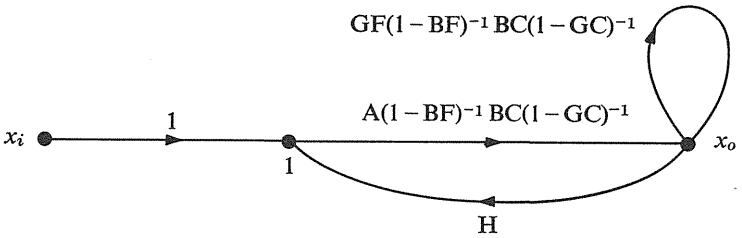


FIGURE 2.29 *A single phase equalizer installed in the transmitter serves all receivers.*

ANSWER TO QUESTION 2.6 Your graph should look like



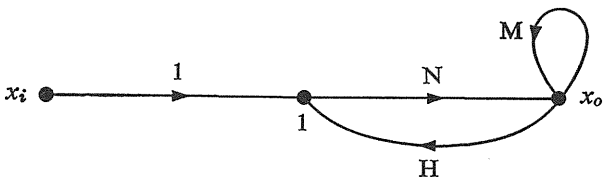
ANSWER TO QUESTION 2.7



These expressions are getting very unwieldy, so let us simplify our work by making the substitution:

$$M = GF(1 - BF)^{-1}BC(1 - GC)^{-1},$$
$$N = A(1 - BF)^{-1}BC(1 - GC)^{-1}.$$

Then



equalizer in each receiver tuned in to that particular station. This possibility is schematized by the flow-graphs in Figures 2.28 and 2.29. In the first graph, the video signal x_1 at the first receiver is given in terms of the video signal x_t at the transmitter by $x_1 = x_t TA_1R_1E$, whereas in the second graph, $x_1 = x_t ETA_1R_1$. However, to the extent that these operators are linear and stationary, the overall result will be the same in the two cases. Hence, we may build into the transmitter a *single* operation that otherwise would have to be performed *many times* at much greater cost.

Considerable algebraic simplification is also achieved when the operators commute. Since order is then not important, an expression like AC/(1 − B) has *explicit* meaning, whereas for noncommuting operators, we would not know whether it meant $(1 - B)^{-1}AC$, $A(1 - B)^{-1}C$, or $AC(1 - B)^{-1}$. Furthermore, it is possible to simplify the fractional forms employing all of the usual algebraic operations. The great algebraic simplification achieved thereby may be illustrated by returning to the graph previously considered in Figure 2.26, and repeated in Figure 2.30. Elimination of the self-loops

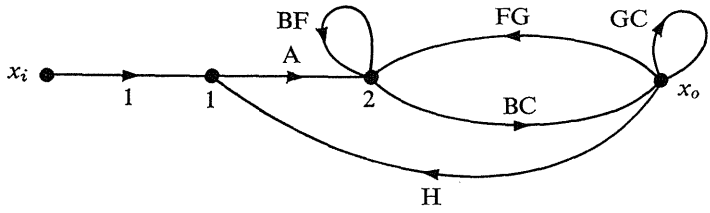
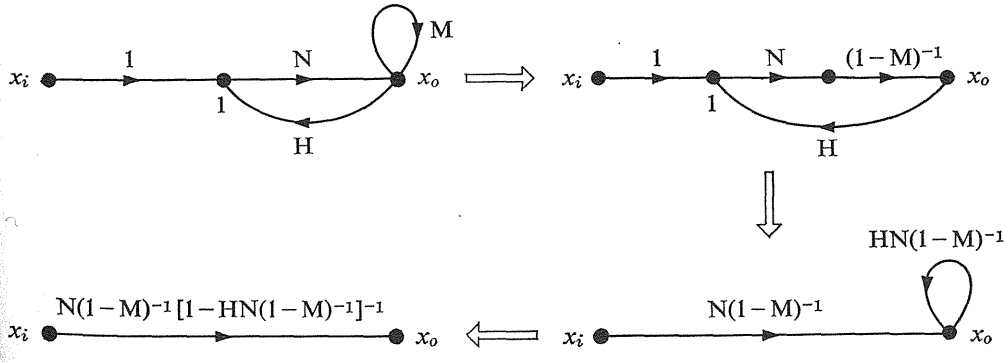


FIGURE 2.30

yields Figure 2.31, where the various products of operators may now be written in any order. Observe that the *self-loop* on each node modifies only the incoming branches to that node.

ANSWER TO QUESTION 2.8



$$K = N(1 - M)^{-1}[1 - HN(1 - M)^{-1}]^{-1}$$

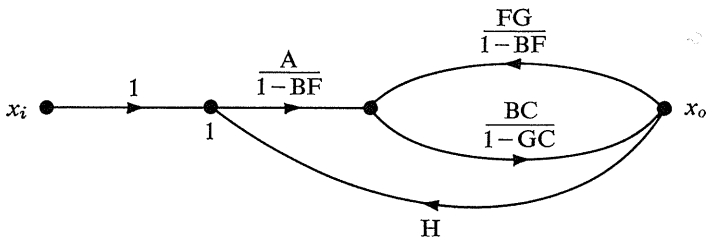


FIGURE 2.31

Next, elimination of node 2 gives

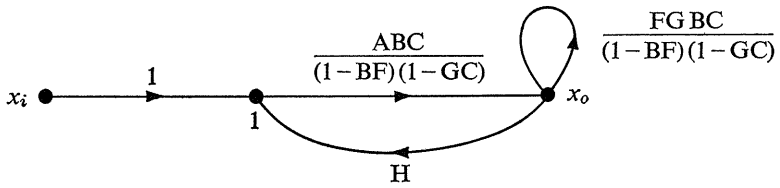


FIGURE 2.32

Elimination of the self-loop yields

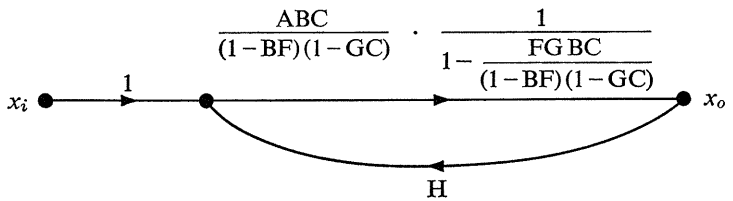


FIGURE 2.33

or, on *multiplying out* the fractional form,

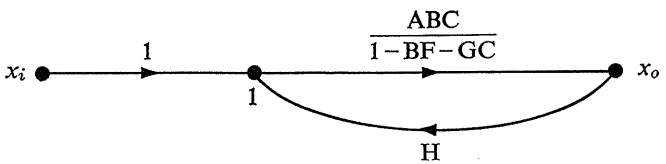


FIGURE 2.34

Observe the great simplification that has occurred as a consequence of being able to treat the operators like ordinary algebraic fractions.

Finally, we may eliminate node 1 to obtain

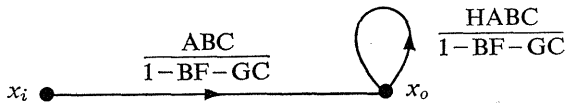


FIGURE 2.35

or

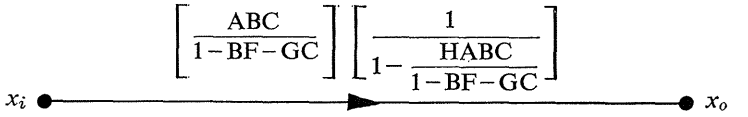


FIGURE 2.36

or

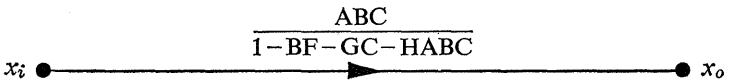


FIGURE 2.37

To summarize: When the operators are *linear* and *stationary* they may be manipulated just like familiar algebraic quantities. Then the flow-graph

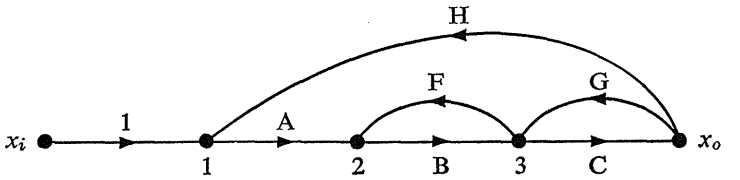


FIGURE 2.38

may be reduced to an equivalent single *graph operator*

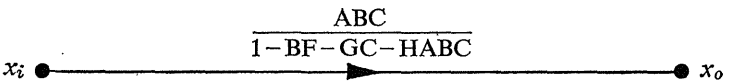


FIGURE 2.39

which is equivalent to writing

$$x_i\left[\frac{\text{ABC}}{1 - \text{BF} - \text{GC} - \text{HABC}}\right] = x_0.$$

To make certain that you fully understand these reduction procedures, next consider a slightly more complicated example. Node 3 is eliminated from the graph shown at the left in Figure 2.40 to obtain the reduced graph shown at the right. All the branch

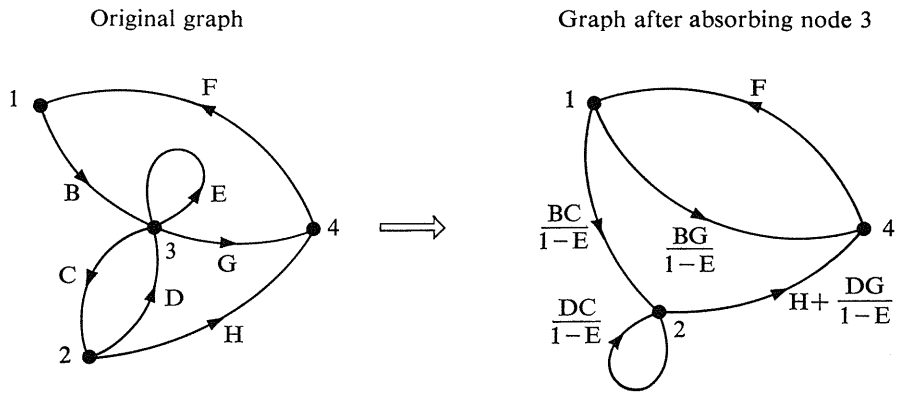


FIGURE 2.40

transmittances of the reduced graph have been expressed in terms of the transmittances of the original graph. Now, carry out your own reduction of the original graph, using the transformation given in Figure 2.17, and by comparing your results with transmittances given above, decide whether the reduction has been correctly made. Also remember that *every* branch that enters a node having a self-loop of transmittance L is modified by the factor $1/(1 - L)$.

The original graph and the reduced graph are equivalent because the transmittances between the retained (i.e., accessible) nodes are identical for the two graphs. For example, the signal at node 1 can travel to node 2 (as well as to other nodes that we will ignore for the present). In traveling this route it is multiplied by B and C branches in cascade, and by $1/(1 - E)$, for the self-loop at 3.

Graph Transmittance (Operator)

It is necessary at this time to define very carefully the different *operator transmittances* that can be associated with a given flow-graph. The transmittance B_{jk} of *any branch jk*

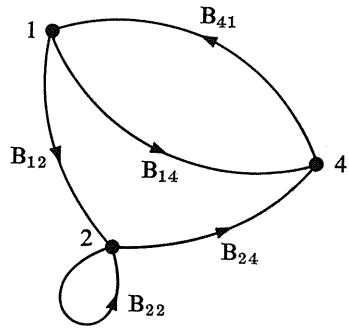


FIGURE 2.41

represents the signal flow into node k per unit of signal at node j with *all other node signals reduced to zero* (i.e., "killed"). These transmittances characterize the relationship between *pairs* of signals, and they may be called *branch transmittances*, to distinguish them from the *graph transmittance*, next to be defined using Figure 2.41.

In the most general case, we shall define the *graph transmittance* G_{jk} *as the signal appearing at node k per unit of external signal injected into node j.* Thus, to find the *graph* transmittance G_{14} from node 1 to node 4, we attach to the graph an external source and an external sink through branches each having unity transmittance, as in Figure 2.42.

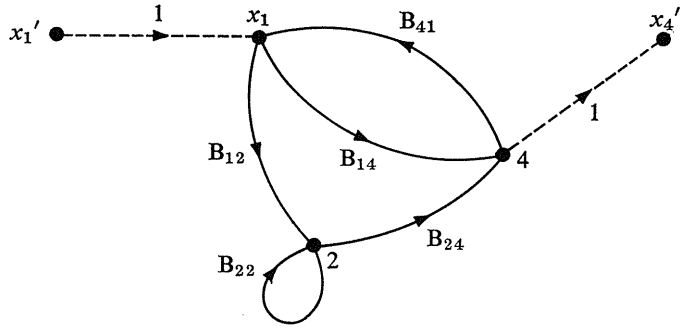


FIGURE 2.42

In Figure 2.43 we have labeled the new source signal with a primed letter x_1' to distinguish it from x_1. The *graph transmittance* G_{14} is then defined as the equivalent transmittance from x_1' to x_4'.

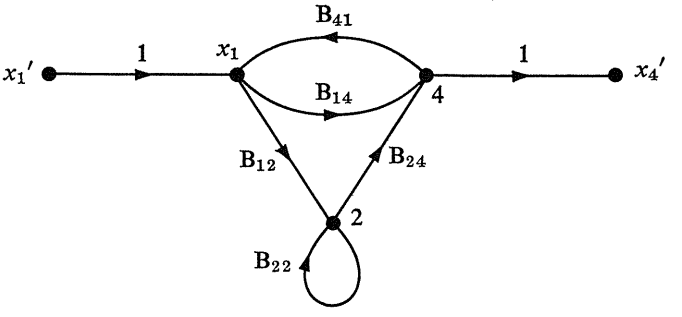


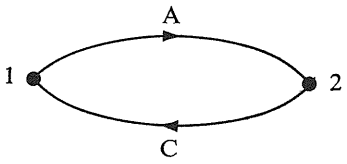
FIGURE 2.43

QUESTION 2.9 Since the transmittance of the added branch from the x_1' node to the x_1 node is unity, is it correct to say that $x_1' = x_1$? (Answer)

QUESTION 2.10 In the same graph, node 4 has been connected to the sink node x_4' via a unity-transmittance branch. Is it correct to assert that the signal at node 4 is identical to the signal x_4'? (Answer)

QUESTION 2.11 Consider the accompanying graph.

1. What is the *branch* transmittance from node 1 to node 2?
2. What is the *graph* transmittance from node 1 to node 2?
3. What is the *branch* transmittance from node 2 to node 1?
4. What is the *graph* transmittance from node 2 to node 1? (Answer)



We have emphasized that the signal at a node depends only on the incoming branch signals. Although the rules for flow-graphs are few in number and easily learned, it is easy to forget some of these definitions. Let us apply the steps illustrated in the answer to Question 2.11 to find the graph transmittance G_{14} of the more complicated graph previously considered. With the "external" source and sink nodes attached as indicated in Figure 2.44, we may then absorb all other nodes, thus reducing the graph to a single branch whose transmittance is G_{14}.

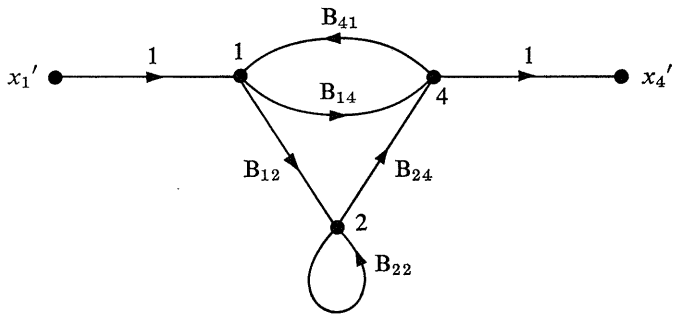


FIGURE 2.44 *"External" source and sink nodes, x_1' and x_4', are introduced to determine the graph transmittance G_{14} from node 1 to node 4.*

In the successive stages of this reduction, illustrated by Figures 2.45, 2.46, and 2.47, node 2 is first absorbed, then node 4, and finally node 1.

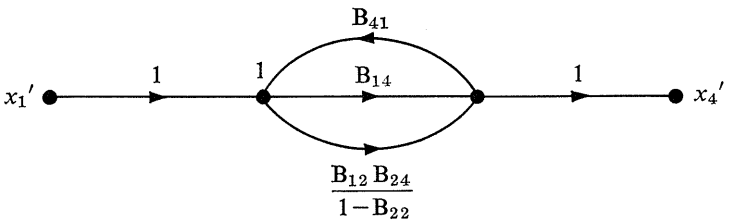


FIGURE 2.45

or

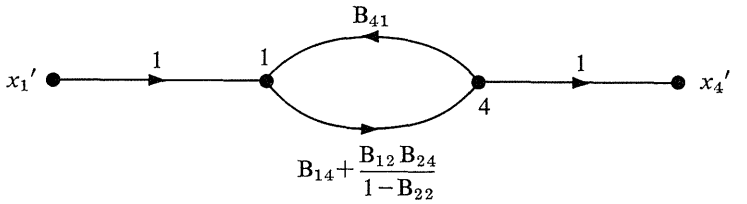


FIGURE 2.46

To simplify the algebra, let

$$C \equiv B_{14} + \frac{B_{12}B_{24}}{1 - B_{22}}$$

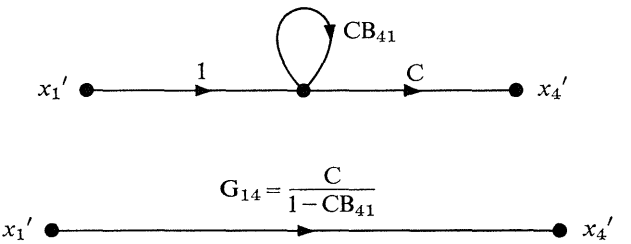


FIGURE 2.47

Expressing this *graph transmittance* in terms of the branch transmittances, we have, with a little reduction,

$$G_{14} = \frac{(1 - B_{22})B_{14} + B_{12}B_{24}}{1 - B_{22} - B_{14}B_{41} - B_{12}B_{24}B_{41} + B_{22}B_{14}B_{41}}.$$

The difference between a *branch transmittance* and a *graph transmittance* should be noted carefully. To get *branch transmittances* only *one* signal was activated at a time,

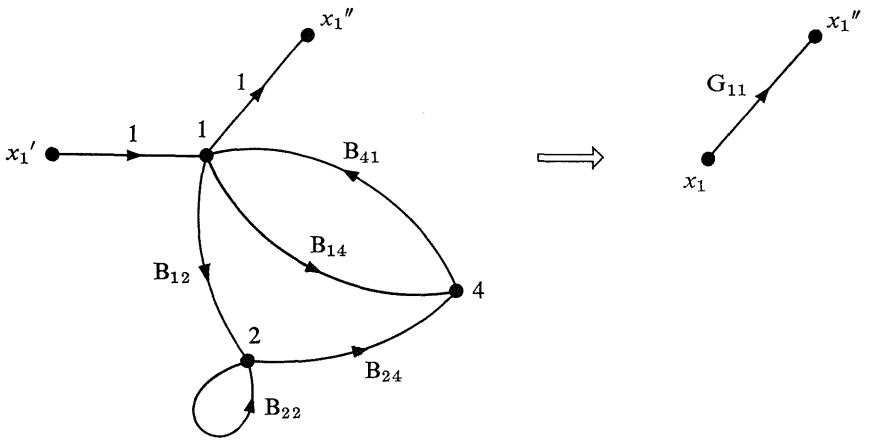


FIGURE 2.48 *The complicated graph at the left may be reduced to a single equivalent branch* G_{11}.

with all other node signals kept at *zero*. For the *graph* transmittance, *all* the node signals were allowed to take on freely their proper value corresponding to signal flow into them. Thus, although the applied source is connected to node 1 over a branch having a transmittance of unity, the *total* signal flow into 1 must include that over B_{41} as well as the signal injected by the source. This defines, for instance, the graph transmittance G_{11} associated with node 1, as shown in Figure 2.48, where, again, by absorbing first node 2 and then node 4 we obtain the single transmittance from which it follows that

$$G_{11} = \frac{1}{1 - CB_{41}}.$$

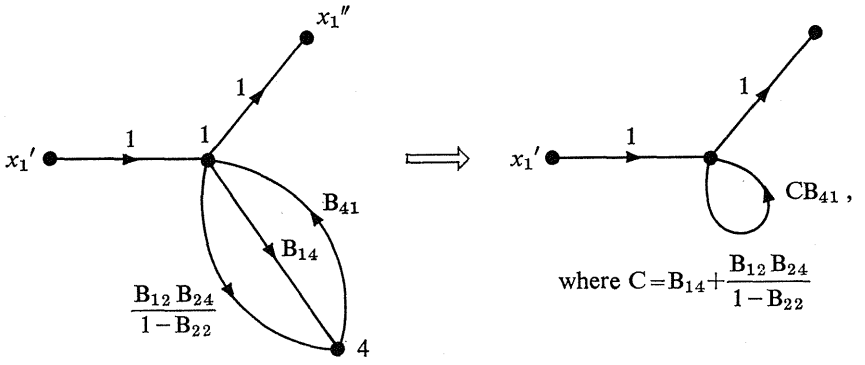
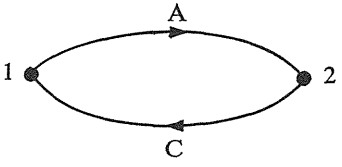


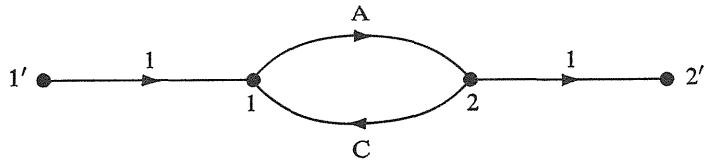
FIGURE 2.49 *Steps in the reduction in the graph of Figure 2.48.*

ANSWERS TO QUESTIONS 2.9, 2.10, 2.11 The distinction between a *branch* transmittance and a *graph* transmittance may be clarified if it is noted that the *branch* transmittance is concerned *only* with paths between two nodes that pass through no other nodes. In contrast, the *graph* transmittance involves *all possible paths* between two nodes (or from a node back to itself).

We may illustrate this distinction by the accompanying very simple graph. Here, A is



the branch transmittance B_{12}, whereas C is the branch transmittance B_{21}. To find the *graph* transmittance G_{12}, we need to find the signal x_2 per unit of signal "injected" into node 1 by an *external source*. Thus, we augment the graph with two unity-transmittance scalor branches, one branch for "injecting" this signal into node 1 and the other



branch for "observing" the signal at node 2. Now the identity operator connecting node 1′ to node 1 will inject the signal x_1' into node 1. However, there is also additional signal flow into node 1 from node 2. Hence, the signal x_1 is *not the same* as the source signal x_1'. In fact,

$$x_1' + x_2\mathrm{C} = x_1.$$

The signal at node 2′, however, is identical to the signal at node 2 because no other signal flows into node 2′. The *graph* transmittance G_{12} describes the relation between the nodes 1′ and 2′. (Even though these "input" and "output" nodes were not explicitly shown on the original graph, they may, of course, always be added to any graph.) To find the signal x_2' we note that

$$x_2 = x_2',$$
$$x_1\mathrm{A} = x_2,$$
$$x_1' + x_2\mathrm{C} = x_1.$$

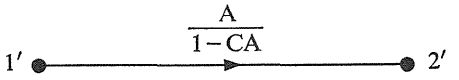
Hence, by substituting the last equation for x_1 into the second equation, we get

$$(x_1' + x_2\mathrm{C})\mathrm{A} = x_2,$$
$$x_1'\mathrm{A} = x_2(1 - \mathrm{CA}),$$

or

$$x_2 = x_1'\mathrm{A}/(1 - \mathrm{CA}),$$

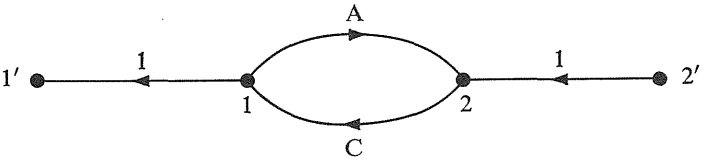
whence



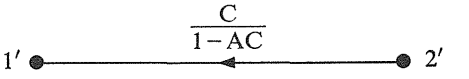
so that

$$\mathrm{G}_{12} = \mathrm{A}/(1 - \mathrm{CA}).$$

In a similar fashion, to obtain the *graph* transmittance from node 2 to node 1, we augment the graph with two input and output nodes and branches:

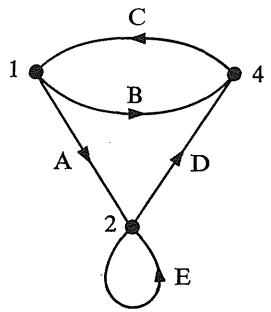


Now $x_2' \neq x_2$, although $x_1 = x_1'$. Reduction of this graph yields



so that $G_{21} = C/(1 - AC)$. It is evident that $G_{12} \neq G_{21}$. Furthermore, there is, in general, *no simple* relation between G_{12} and G_{21} (except for their relation implied by the graph itself). For instance, G_{21} is *not* the reciprocal of G_{12}. In Chapter 1, we talked about Jones' automobile in which adding gas to the tank (i.e., signal injection into x_1) produced a deflection x_2 of the gas gage. (This relation is expressed by the graph transmittance G_{12}.) We agreed that to produce an extra deflection of the gage by means of a small magnet (i.e., signal injection into x_2) would not result in a change in the gas tank x_1. (This relation is expressed by the graph transmittance G_{21}.) It is clear that G_{12} is certainly unrelated to G_{21} in the case of Jones' car, and they are also unrelated in general.

QUESTION 2.12 Now to get some practice, what is the graph transmittance G_{12} associated with the graph shown? (Answer)



QUESTION 2.13 What are the graph transmittances G_{21}, G_{24}, G_{42}, and G_{41}? Express each one as a simple algebraic fraction in the operators A, B, C, D, and E. Also show that $G_{21} \neq G_{12}$. (Answer)

The Graph Determinant

Perhaps you have already noted that the numerator of each of these expressions contains the transmittance of the open path leading from the source to the sink in each case. Surely you also must have already observed that the denominators of these fractions are all identical. If you will calculate the values of the graph transmittances G_{11}, G_{22}, and G_{44}, you should be even more convinced that the quantity $(1 - E - BC - ADC + BCE)$ must be very intimately associated with this particular flow-graph, for it appears in the expression for every graph transmittance. The situation here resembles that encountered in solving a set of simultaneous algebraic equations by means of determinants. You will recall that the determinant formed from the coefficients of the equations appears in the denominator of the solution for each of the variables. Because of this similarity, we shall define the quantity which appears as the denominator of every graph transmittance as the *graph determinant*. We shall now consider the graph determinant in some detail.

First, let us see if the various terms in the graph determinant can be related to any obvious features of the original flow-graph.

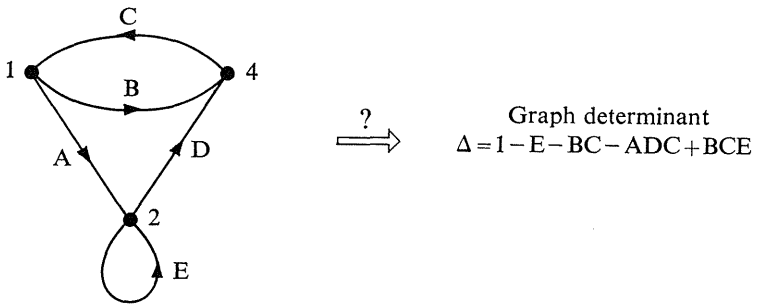
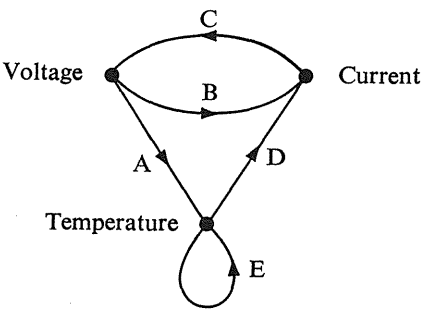


FIGURE 2.50 *Can the graph determinant be found by direct inspection of the graph?*

The determinant is the sum of several terms. The first term is 1 (or, more generally, the *identity* operator). It is *dimensionless.* The dimension of every term, such as ADC, in the determinant must then be the same as 1—that is to say, the individual terms of the graph determinant are *all* dimensionless. This is an interesting and important result, for dimensionless quantities do not depend on the particular physical units that may be used to measure the observables. Hence, these particular groups of operations may be expected to reveal important *invariant properties* of the system represented by the flow-graph. As we shall see, this is actually the case.

To see what other important consequences follow from the dimensionless nature of the graph determinant, let us assume for purposes of illustration that x_1 is a *voltage*, x_2 is a *temperature*, and x_4 is a *current*. Then, C is an operator that converts a current into a voltage, and hence may be identified as a *resistance* (or more generally, as an *impedance*). Likewise, B is an operator that converts a voltage into a current, and it may be identified as a *conductance* (or more generally, as an *admittance*). The operator A evidently converts a voltage into a temperature; E a temperature into a temperature; and D a temperature into a current.

QUESTION 2.14 If the composite operator ADC is applied to a voltage, what kind of physical signal will result? (Answer)

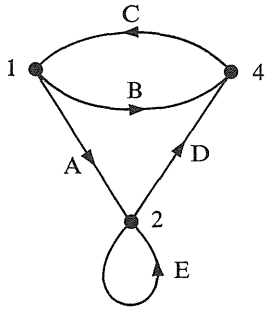


QUESTION 2.15 If the composite operator DCA is applied to a temperature, what kind of physical signal will result? (Answer)

QUESTION 2.16 The sequences of operations E, BC, and ADC each describes a path through the graph. What is similar about these paths? (Answer)

The possibility of associating loops with the additive terms of the graph determinant looks promising. But how can we account for a term such as BCE? This term is evidently the product of *two loops*, (BC) and (E). Notice that there are *other* possible products of two loops, such as (ADC) (BC), and (ADC) (E), and these products do *not* appear in the graph determinant.

QUESTION 2.17 Examine carefully the pairs of loops described by



(BC) (E)
(ADC) (BC)
(ADC) (E)

What unique property distinguishes the particular loop-pair (BC)(E) from the other two loop-pairs? (Answer)

The possibility that the graph determinant may be the sum of loop transmittances plus the product of *nontouching* loops is an appealing hypothesis. We notice that the term BCE has a plus sign, whereas E, BC, and ADC each carries a minus sign. But note that

$$\text{BCE} = (-\text{BC})(-\text{E}),$$

so BCE with a plus sign is actually consistent with the first three terms.

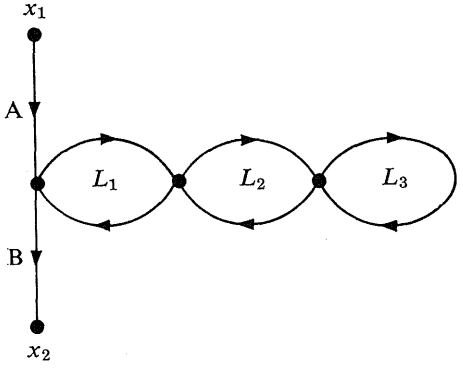
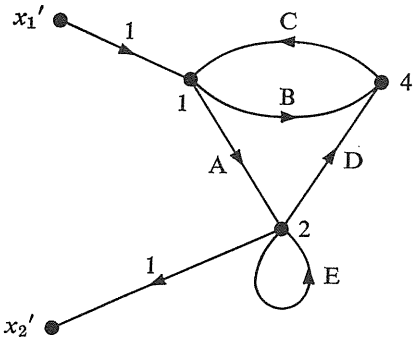


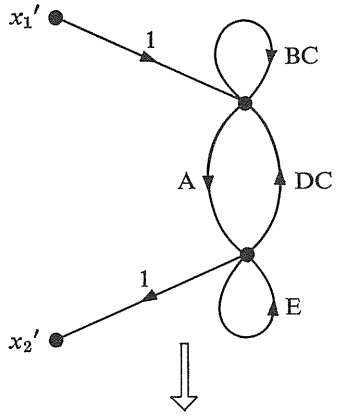
FIGURE 2.51

ANSWER TO QUESTION 2.12 The reduction of this graph to find the graph transmittance G_{12} is accomplished by first attaching source and sink nodes through unity transmittances and then eliminating all other nodes just as in the previous example. The following diagrams should be self-explanatory.

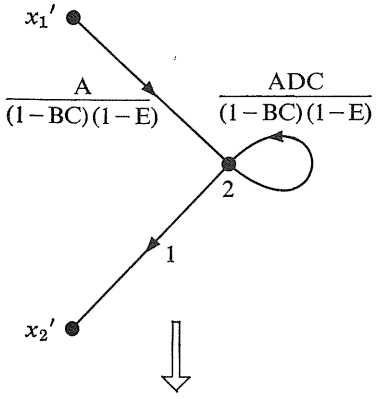
Attach "input" and "output" branches



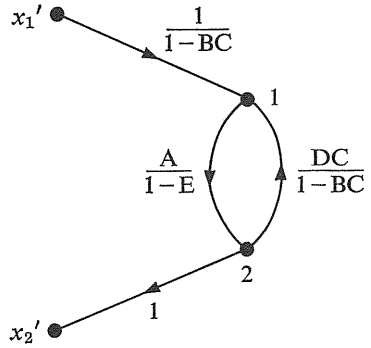
Eliminate node 4



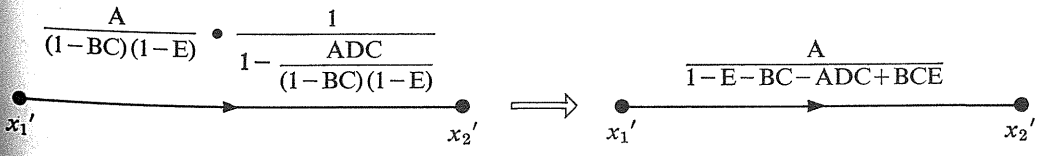
Eliminate node 1



Eliminate self-loop



Eliminate node 2



A quick way to check the validity of a hypothesis is to see if it accounts for other known results. Let us find the graph transmittance from 1 to 2 of the graph in Figure 2.51 (p. 82) by the reduction methods already developed. Here the loop transmittances are L_1, L_2, and L_3, respectively.

Since 1 and 2 are already *source* and *sink nodes*, respectively, there is no need for adding extra branches to this graph. We may use the loop-elimination rule to achieve an immediate reduction:

$$G_{12} = \frac{AB}{1 - \dfrac{L_1}{1 - \dfrac{L_2}{1 - L_3}}} \equiv \frac{(1 - L_2 - L_3)AB}{1 - L_1 - L_2 - L_3 + L_1L_3}.$$

QUESTION 2.18 Does our hypothesis that the graph determinant is composed of 1 plus the sum of the negative transmittances of loops and the products of nontouching pairs hold true for the graph in Figure 2.51 ? (Answer)

QUESTION 2.19 The numerator of the expression just found $(1 - L_2 - L_3)AB$ also involves various products. Why do you suppose L_2 and L_3 appear in combination with AB, whereas L_1 is missing? (What geometrical property does L_1 have that neither L_2 nor L_3 possess relative to AB?) (Answer)

QUESTION 2.20 Recall that we wish to find the graph transmittance from 1 to 2. Does the operator group AB appearing in the numerator appear to have any relation to the source and sink nodes? State the relationship. (Answer)

ANSWER TO QUESTION 2.13

$$G_{21} = \frac{DC}{1 - E - BC - ADC + BCE}$$

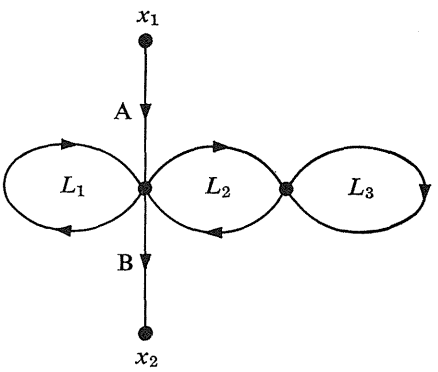
$$G_{24} = \frac{D}{1 - E - BC - ADC + BCE}$$

$$G_{42} = \frac{CA}{1 - E - BC - ADC + BCE}$$

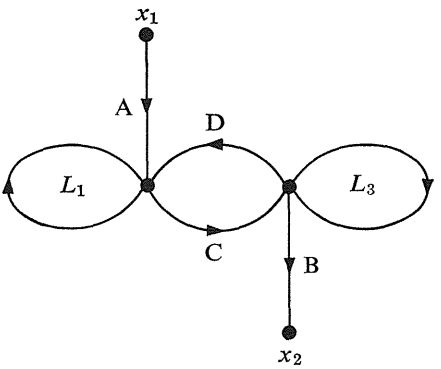
$$G_{41} = \frac{C(1 - E)}{1 - E - BC - ADC + BCE}$$

ANSWER TO QUESTIONS 2.14, 2.15, 2.16 It should be clear that the paths E, BC, and ADC each corresponds to a *loop* of the graph. (Remember, a *loop* is a closed path in which no node is traversed more than once.) This is not surprising, since these composite operator sequences are dimensionless and must therefore yield as their "outputs" the same kind of observable as their "inputs." They begin and end on the same node. A node does not operate on a signal.

QUESTION 2.21 Try out any hypotheses that you may have formed about the numerator of the graph transmittance by evaluating the graph transmittance of the accompanying graph (L_1, L_2, and L_3 are loop transmittances). What are the numerator and denominator expressions in this case? (Answer)



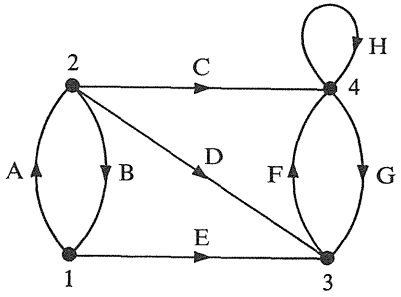
QUESTION 2.22 Furthermore, verify and refine your hypotheses about the formation of the numerator of the graph determinant by evaluating the graph transmittance for the accompanying graph, where the loop transmittance $L_2 = DC$. (Answer)



ANSWER TO QUESTION 2.17 One property of the loop-pair (BC)(E) is that the two loops are nontouching. That is, *loop* BC *has no node in common with loop* E. In contrast, the other two pairs have one or more nodes in common. Hence, ADC and BC are *touching* loops, as are also ADC and E.

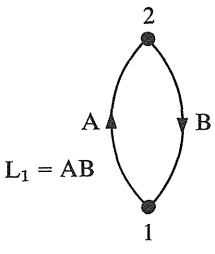
ANSWER TO QUESTION 2.18 To state the rule for forming the graph determinant, which we shall denote by Δ, consider the graph shown. The determinant, Δ, of this graph is given by

$$\Delta = 1 - (AB + FG + H) + (ABFG + ABH).$$

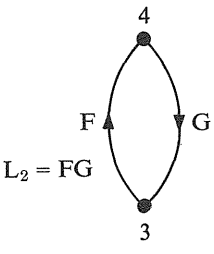


The determinant evidently may be expressed as unity plus (or minus) various products of the branch transmittances. Several interesting properties may be noted on careful observation.

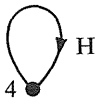
1. Only transmittances of branches which form a loop appear in the determinant. Branches C, D, and E, which are *not* part of any loop, do *not* appear.
2. The successive terms in the first parentheses are the loop transmittances L_k of the three loops contained in this graph:



$L_1 = AB$



$L_2 = FG$



$L_3 = H$

Furthermore, the successive terms in the second parentheses are all the possible transmittance products of *pairs of nontouching* loops:

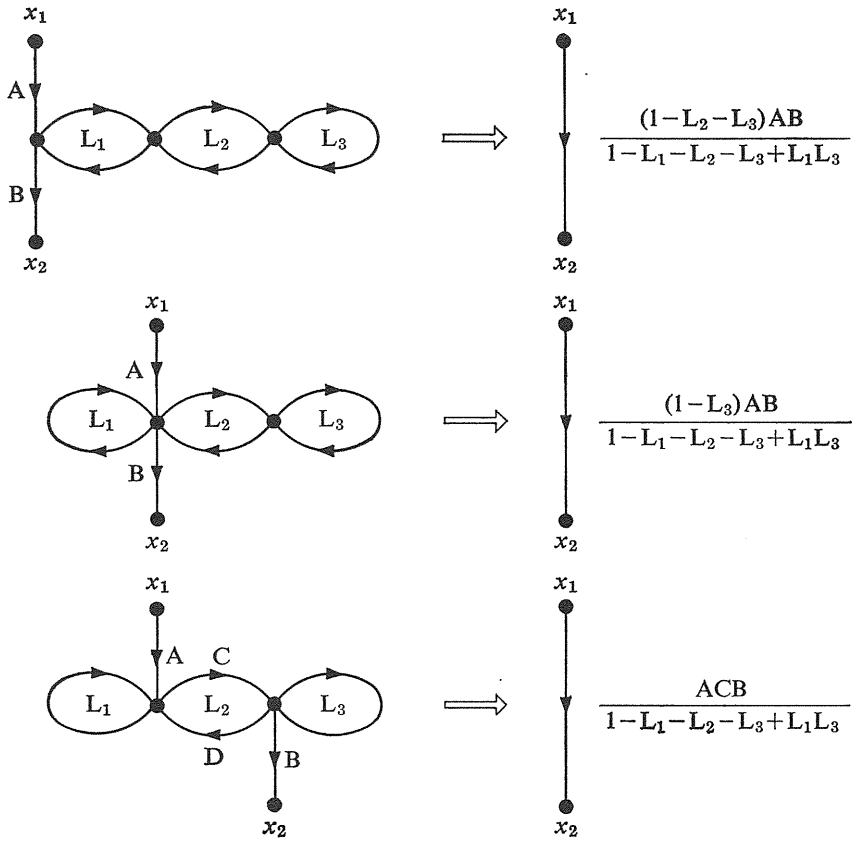
$$\Delta = 1 - (L_1 + L_2 + L_3) + (L_1L_2 + L_1L_3).$$

This expression would be identical to the value of

$$\Delta = (1 - L_1)(1 - L_2)(1 - L_3)$$

provided that we agree to delete all terms containing products of touching loops. In this example, L_2L_3 is the only term involving touching loops.

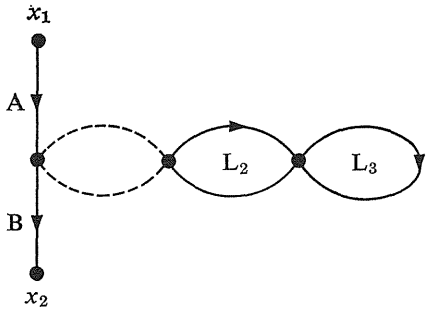
ANSWERS TO QUESTIONS 2.19, 2.20, 2.21, 2.22 The graph transmittances of the three graphs considered in these questions are tabulated here:



From this, it is evident that the graph determinant is the *same* in each case, since it depends only upon the *loop graph* (i.e., the graph which is left after all branches that are not part of some loop have been deleted). The numerator is, however, *different.*

In each of these cases, the numerator involves the *product of branches which form an open path from the specified source node to the specified sink node.* In the third graph, this transmittance ACB of this open path occurs alone in the numerator. But in the first two graphs, the path transmittance AB is multiplied by a factor that looks suspiciously like a *graph determinant.*

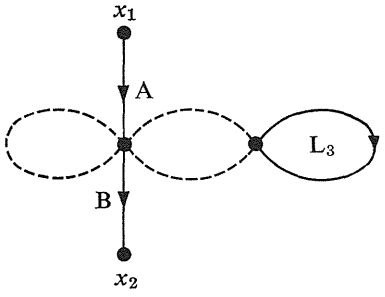
The factor $(1 - L_2 - L_3)$ associated with the path AB in the first graph, we shall call a *path cofactor.* It would be the determinant of a graph that contained only two *touching* loops L_2 and L_3. Such a graph would result if we were to *delete* all parts of the original graph that *touched* any part of the open path AB from source to sink.



Path transmittance is AB.
Graph determinant is $1 - L_2 - L_3$.

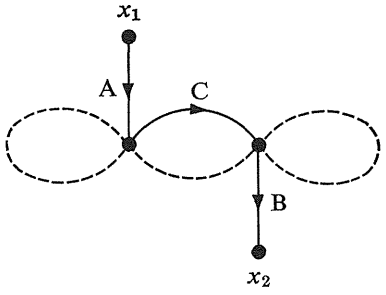
This seems like a beautifully simple idea! Perhaps it will apply generally. Let us see.

For the second graph, we delete all parts that touch the open path AB from source to sink, which again yields the correct answer.



Path transmittance is AB.
Graph determinant is $1 - L_3$.

Let us try the procedure on the third graph. Here, *all* loops are *destroyed* so that



Path transmittance is ACB.
Graph determinant is 1.

These three examples certainly support the hypothesis that the graph transmittance between a specified *source* and a specified *dependent* node is given by a rational fraction in the operators composing the graph. The denominator of this rational fraction is simply the graph determinant, Δ. The numerator is simply the transmittance of the open path (from source to dependent node) multiplied by the *graph* determinant of a *new graph* formed by temporarily *deleting* from the original graph any branches *that touch the open path.*

That the graph transmittance could be formed this way was demonstrated by Samuel Mason, who first invented signal flow-graphs, in his doctoral dissertation written at Massachusetts Institute of Technology in 1951 (where he is now a professor, working on artificial sensory aids for sensory deprived people). In Appendix A at the end of this chapter, we shall outline Mason's proof of this relationship as given in Reference 3 in Appendix B.

Mason's Loop-Expansion Theorem

We have just discovered a major result first given by Mason (see Appendix A at the end of this chapter) which permits us to expand the graph determinant directly in terms of the loop transmittances.

Mason's Loop-Expansion Theorem
For any graph having m different loops, the determinant may be written as

$$\Delta = [(1 - L_1)(1 - L_2) \ldots (L - L_m)]^*, \tag{2.4}$$

where * indicates that *all* terms containing products of *touching* loops are to be deleted.
An equivalent form is

$$\Delta = \left[1 - \sum_k L_k + \sum_{j,k} L_j L_k - \sum_{j,k,v} L_j L_k L_v + \cdots\right]^*, \tag{2.5}$$

$$k > j > v$$

where the * again indicates that all terms containing products of touching loops are to be deleted.

To apply Mason's expansion theorem, we must first recognize *all* the loops belonging to a given graph. This task is often simplified by drawing the *loop subgraph*, which results when all branches that are not part of some loop are removed. For the example just considered, the loop subgraph is obtained by removing the branches C, D, and E from the original graph. When the loop subgraph consists of two or more nontouching

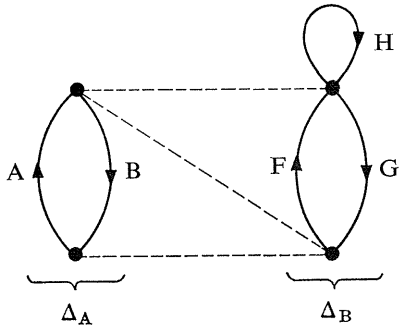


FIGURE 2.52

parts, it follows from Mason's expansion theorem that *the determinant of the complete flow-graph is equal to the product of the determinants of each of the nontouching parts of the loop subgraph.* In this example,

$$\Delta_A = 1 - AB = 1 - L_1,$$
$$\Delta_B = 1 - FG - H = 1 - L_2 - L_3.$$

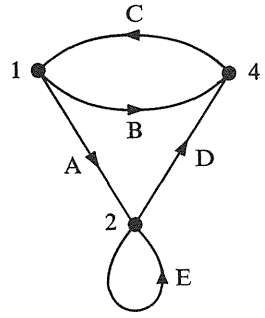
Hence

$$\begin{aligned}\Delta &= \Delta_A \cdot \Delta_B \\ &= (1 - L_1)(1 - L_2 - L_3) \\ &= 1 - (L_1 + L_2 + L_3) + (L_1L_2 + L_1L_3),\end{aligned}$$

just as before.

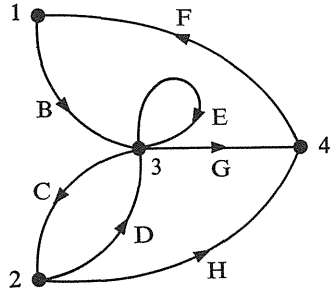
We have now nearly reached our goal. The ability to evaluate the graph determinant by simple inspection is so essential to what follows that before proceeding further it will be wise to work several examples to make certain that you have mastered these points.

QUESTION 2.23 Consider the accompanying graph.



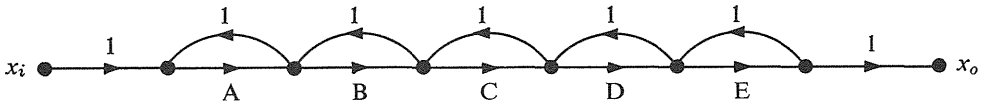
1. How many branches are *not* part of a loop?
2. How many different loops are there?
3. How many different nontouching pairs of loops are there?
4. How many different nontouching triplets of loops are there?
5. Determine the graph determinant using Mason's loop-expansion theorem (Answer)

QUESTION 2.24 For the accompanying graph:



1. How many branches are *not* part of a loop?
2. How many different loops are there? Draw them.
3. How many different nontouching pairs of loops are there?
4. How many different nontouching triplets of loops are there?
5. Determine the graph determinant using Mason's loop-expansion theorem. (Answer)

QUESTION 2.25 For the accompanying graph:



1. How many branches are *not* part of a loop?
2. How many different loops are there?
3. How many different nontouching pairs of loops are there? Label the loops and tabulate your answer.
4. How many different nontouching triplets of loops are there?
5. Determine the graph determinant using Mason's loop-expansion theorem. (Answer)

In the graphs just considered there was only *one* open path, from *source node* to *dependent node*. In general, there may be several paths. How does this affect the conclusions that we just reached?

Again, let us explore the possibilities by considering a specific case, such as Figure 2.53.

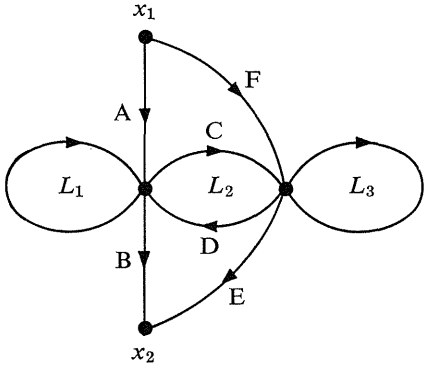


FIGURE 2.53

In this graph, we now have *four different open paths* from source to sink. We shall first reduce this graph, using the methods previously developed.

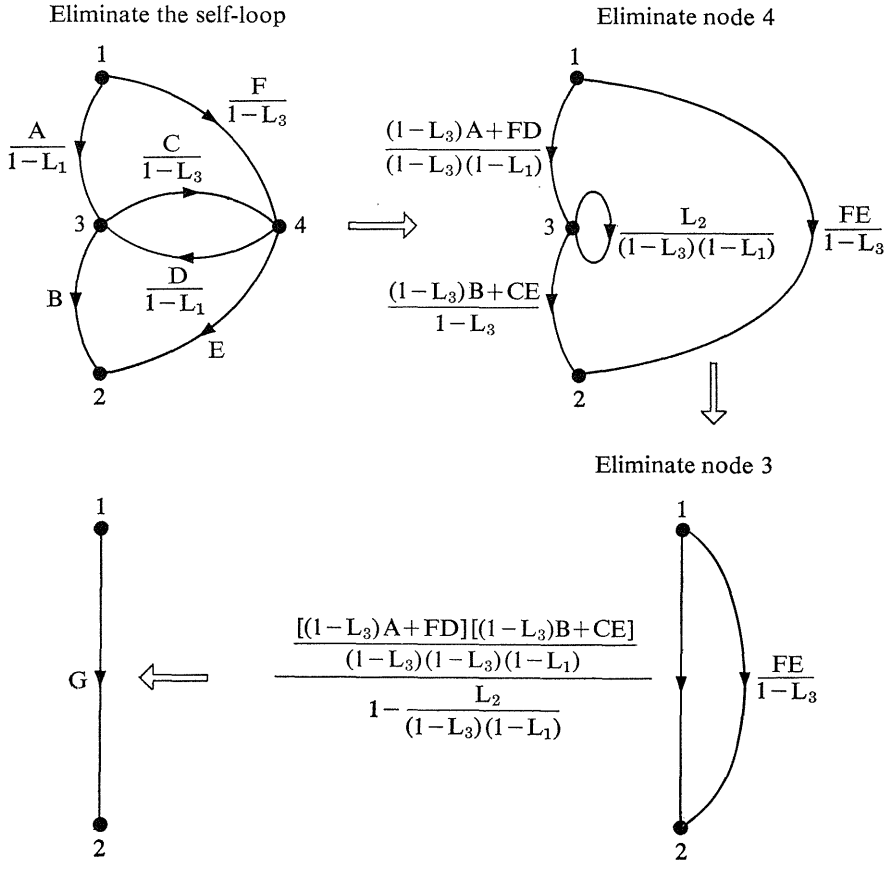


FIGURE 2.54 *Steps in the reduction of the graph of Figure 2.53:*

1. *Eliminate the self loop;*
2. *Eliminate node 4.*
3. *Eliminate node 3 to obtain graph 4.*

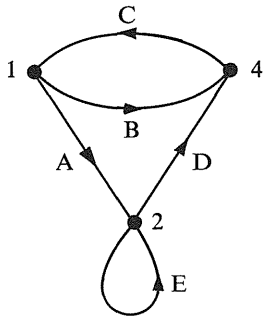
Where

$$G = \frac{[(1 - L_3)(1 - L_1) - L_2]FE + [(1 - L_3)A + FD][(1 - L_3)B + CE]}{[1 - L_3][(1 - L_3)(1 - L_1) - L_2]}.$$

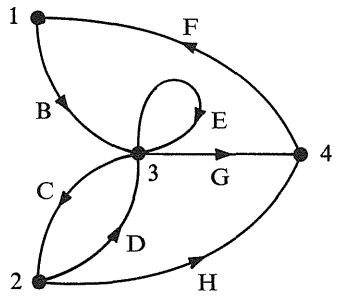
Further reduction of the rather messy expression for G is possible by noting that, since $L_2 = DC$ by definition, two of the terms in the numerator will cancel when the expression is multiplied out (that is, $-L_2FE$ and FDCE cancel to zero). Furthermore, it is then possible to divide out a common factor, $(1 - L_3)$, from both numerator and denominator. This finally yields

$$G = \frac{(1 - L_3)AB + (1 - L_1)FE + ACE + FDB}{1 - L_1 - L_2 - L_3 + L_1L_3}.$$

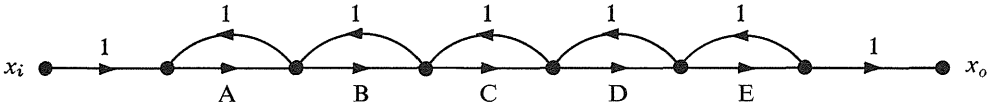
ANSWERS TO QUESTIONS 2.23, 2.24, 2.25



Graph in Question 2.23



Graph in Question 2.24



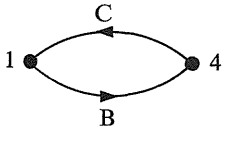
Graph in Question 2.25

1. How many branches are *not* part of some loop?

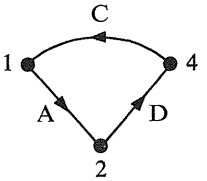
For the graph in Question 2.23

2. How many different loops are there?

Graph in Question 2.23



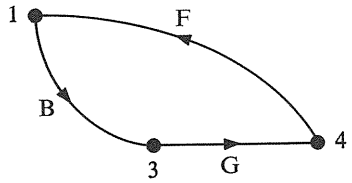
$L_1 = CB$



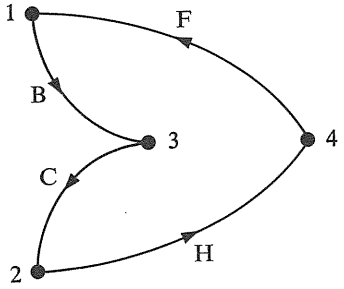
$L_2 = CAD$

$L_3 = E$

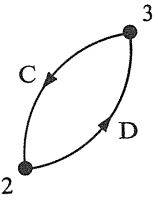
Graph in Question 2.24



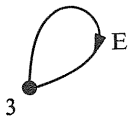
$L_1 = BFG$



$L_2 = CHFB$

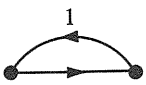


$L_4 = CD$

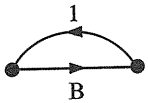


$L_3 = E$

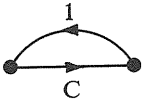
Graph in Question 2.25



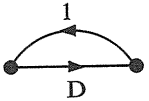
$L_1 = A$



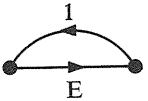
$L_2 = B$



$L_3 = C$



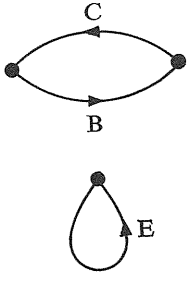
$L_4 = D$



$L_5 = E$

3. How many different nontouching *pairs* of loops are there?

Graph in Question 2.23



One *pair*: L_1, L_3

Graph in Question 2.24: All pairs touch. Hence, there are *no* nontouching pairs of loops.

Graph in Question 2.25: Six pairs of nontouching loops.

$$L_1L_3,\quad L_1L_4,\quad L_1L_5,$$
$$L_2L_4,\quad L_2L_5,$$
$$L_3L_5.$$

4. How many different nontouching *triplets* of loops are there?

   Graph Q.2.23: None, all triplets touch.

   Graph Q.2.24: None, all triplets touch.

   Graph Q.2.25: One, $L_1L_3L_5$.

5. Determine the graph determinant using Mason's loop-expansion theorem. By using the answers of questions 2, 3, and 4 in

$$\Delta = \left[1 - \sum_k L_k + \sum_{j,k} L_jL_k - \sum_{i,j,k} L_iL_jL_k + \cdots\right]^*,$$

   we get

   Graph Q.2.23: $\Delta = 1 - (L_1 + L_2 + L_3) + (L_1L_3)$,

   Graph Q.2.24: $\Delta = 1 - (L_1 + L_2 + L_3 + L_4)$,

   Graph Q.2.25: $\Delta = 1 - (L_1 + L_2 + L_3 + L_4 + L_5)$
   $+(L_1L_3 + L_1L_4 + L_1L_5 + L_2L_4 + L_2L_5 + L_3L_5)$
   $-(L_1L_3L_5)$.

Here, our graph determinant emerges just as before. Furthermore, the numerator is seen to consist of the sum of *four* groups of terms, each of which is comprised of the path transmittance of one of the four open paths (AB, FE, ACE, and ADB), multiplied by the determinant of the graph remaining after all branches touching that particular path have been deleted. We have thus discovered:

Mason's General Graph Transmittance Expression

Let

G = Source-to-sink graph transmittance (the sink signal produced for each unit of source signal).

P_ν = Transmittance of the νth source-to-sink *open* path.

Δ = Determinant of the original graph.

Δ_ν = Cofactor of the νth path (the determinant of that part of the graph *not touching* the νth path).

and defined two branches to be *confluent* if they *both originate or both terminate* at the same node.

Then, we have the important result:

Mason's General Graph-Transmittance Expression

$$G = \frac{\sum_v P_v \Delta_v}{\Delta}$$

$$= \left[\frac{(P_1 + P_2 + \cdots + P_k)(1 - L_1)(1 - L_2)\ldots(1 - L_m)}{(1 - L_1)(1 - L_2)\ldots(1 - L_m)}\right]^*, \tag{2.6}$$

where * indicates that any terms containing transmittance products of confluent branches are to be dropped.

The cofactor Δ_v is formed by deleting from Δ any terms which contain branches touching the path P_v. This cofactor may also be found by first erasing all parts of the original graph that touch the path P_v and then evaluating the determinant of the subgraph that remains. Either method will yield the same result.

Let us now use Mason's expansion to find the source-to-sink transmittance of Figure 2.55. The determinant Δ of this graph is readily found by the method already described. It remains to find the various open paths that lead from x_s to x_4.

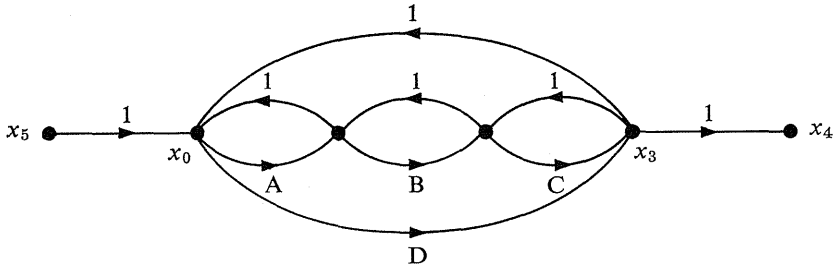


FIGURE 2.55

Examination of the graph shown above reveals that there are only two open paths, one having the transmittance $P_1 = D$, and the second having the transmittance $P_2 = ABC$. Furthermore, the subgraph left after deleting all branches that touch $P_1 = D$ consists of only the center loop B, so that the cofactor Δ_1 of this path is

$$\Delta_1 = 1 - B.$$

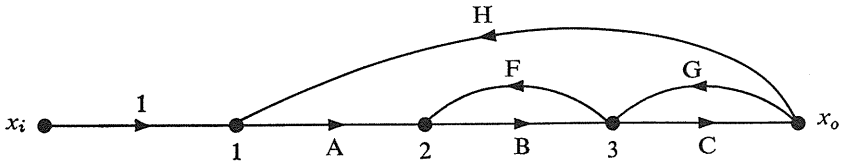
All branches of the graph touch the second path, so that no loops at all are left in its subgraph and its cofactor Δ_2 is 1. Hence,

$$G_{54} = \frac{P_1\Delta_1 + P_2\Delta_2}{\Delta}$$

$$= \frac{D(1 - B) + ABC}{1 - (A + B + C + 2D + ABC) + (AC + BD)}.$$

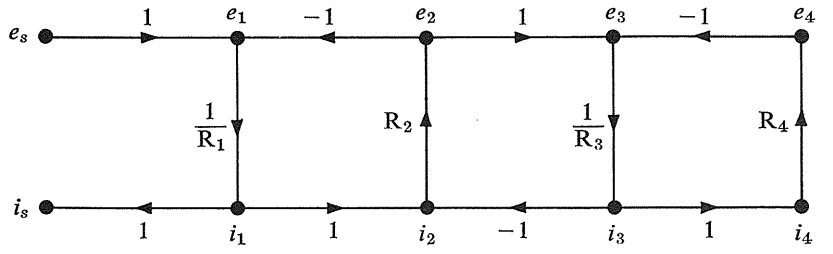
Thus, by means of Mason's transmittance expression, we can write down *by inspection* the graph transmittance between any source node and dependent node.

QUESTION 2.26 Consider the flow-graph for the hi-fi amplifier (redrawn from Figure 2.24 of this chapter):



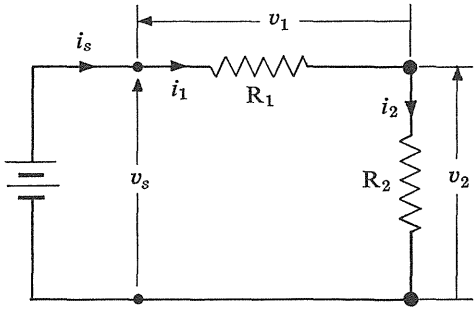
How much more quickly can you now evaluate the graph transmittance from x_i to x_o using Mason's method than when using the first method described for solving this same problem? (See pp. 67–68.) What are the explicit expressions for the graph transmittances from x_i to node 3; to node 2; to node 1? (Answer)

QUESTION 2.27 In the flow-graph shown herewith, e_s is a source signal and all other signals depend on e_s.



1. How many loops are there in this graph? Label these loops L_1, L_2, and so forth, and express the loop transmittance in terms of the transmittances of the individual branches.
2. What is an expression for the equivalent graph transmittance G that expresses i_s directly in terms of e_s and the loop transmittances found in part A?
3. Suppose that the parameters denote *resistances*. Their values are given as $R_1 = 2$ ohms, $R_2 = 2$ ohms, $R_3 = 1$ ohm, and $R_4 = 3$ ohms. What will be the value of the loop transmittances L_1, L_2, L_3? What are the units in terms of which the loop transmittances are expressed?
4. If *all* the resistance values were multiplied by a constant factor of 377, by what factor would the *loop* transmittances be multiplied? By what factor would the *path* transmittance be multiplied?
5. What is the value of the graph transmittance for the particular parameters given in part 3? (Answer)

QUESTION 2.28 A common electric circuit is the *resistance voltage divider*, as shown. In this circuit, a source of voltage, such as a battery, produces a specified voltage v_s. As a consequence, currents flow *through* the two resistances. The pertinent relations are:



Among the voltages: The voltage v_1 *across* the resistor R_1 when added to the voltage v_2 *across* the resistor R_2 is equal to the voltage v_s developed by the battery. Hence,

$$v_1 + v_2 = v_s \qquad \text{(Kirchhoff's voltage law).}$$

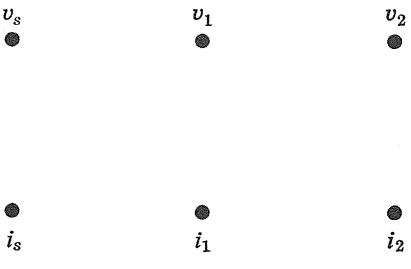
Among the currents: The three elements are connected in *series* so that the current *through* each element is equal to the currents *through* the other two. That is,

$$i_1 = i_s, \qquad i_1 = i_2 \qquad \text{(Kirchhoff's current law).}$$

Between the voltages and currents: The voltage across an ideal resistor is directly proportional to the current through it. The coefficient of proportionality defines the *resistance* parameter, such as R_1 and R_2. Hence,

$$i_1R_1 = v_1, \qquad i_2R_2 = v_2 \qquad \text{(Ohm's law).}$$

Consider now the flow-graph involving the six signals that are associated with this three-element circuit (there are one voltage and one current for each element). Remember that voltage is measured between pairs of terminals. It is like geographical elevation in the sense that if the voltage at terminal A is v_1 relative to terminal B, and the voltage at terminal B is v_2 relative to terminal C, then the voltage of terminal A relative to terminal C is $v_1 + v_2$. That is, we can measure the voltage difference between two terminals at remotely separated points in a circuit, even if there is no single circuit element directly connecting them. For instance, in the voltage divider shown, v_s is definable in terms of v_1 and v_2 regardless of the source that is fastened between these terminals.



1. Considering only the relation between the voltages, draw in the branches that will correctly express v_1 in terms of v_s and v_2. (The answer is unique.)
2. Considering only the relation between the currents, draw in the branches that will correctly express i_s and i_2 in terms of i_1. (The correct answer is unique.)
3. Considering only the current-voltage relations, draw in the necessary branches so that with the exception of the source v_s, every signal is expressed in terms of the other signals in the graph. (That is, all signals are dependent on the source signal v_s.)
4. Use Mason's method to express v_2 directly in terms of v_s and the circuit parameters. (That is, what is the graph transmittance from v_s to v_2 ?) Also, express v_1 in terms of v_s (by finding the graph transmittance from v_s to v_1).
5. Using Mason's method, express i_s directly in terms of v_s. What is the value of a single resistance that if connected across the battery would exhibit this same relation between the current i_s and voltage v_s at its terminals? (Answer)

QUESTION 2.29 The discussion of the previous problem revealed that it is sometimes necessary to *rearrange* the relationships between the signals as originally presented if one is to represent the entire set of relations by a single signal flow-graph. A good technique for doing this is illustrated by the following problem.

Given the set of equations

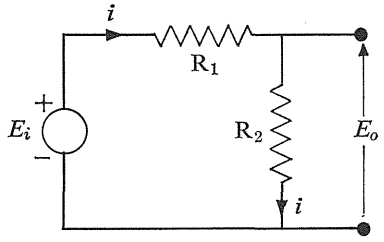
$$6 = x_1 + x_2 + x_3,$$

$$0 = -x_1 + 2x_2 - x_3,$$

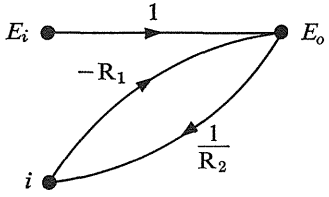
$$0 = -3x_2 + 2x_3.$$

By first suitably rearranging these equations so that there is *one and only one* equation expressing explicitly *each* of the variables in terms of the variables and a constant term (which is easily obtained by introducing a *unit-source* node which is *defined* to have a signal whose value at any instant is $+1$), construct a signal flow-graph representing these rearranged equations. Using the unit node as a source having a constant value of unity, the nodes x_1, x_2, and x_3 then may be expressed directly in terms of this source. Using Mason's rule, solve the graph for the values of x_1, x_2, and x_3. (Answer)

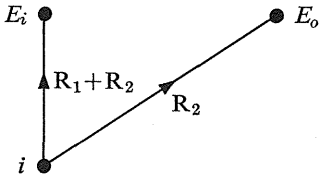
QUESTION 2.30 To see if you have developed any familiarity with the significance of voltage and current in a circuit, determine which of the flow-graphs shown are



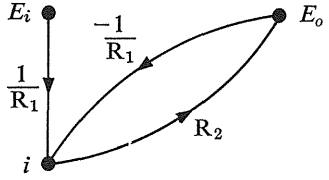
(1)



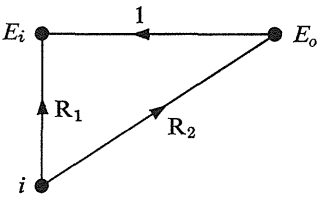
(3)



(2)

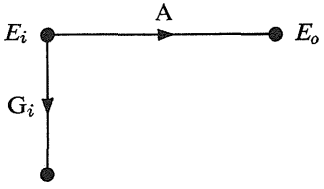


(4)

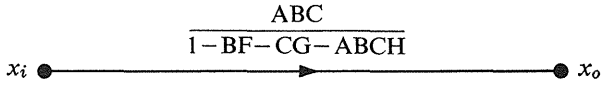


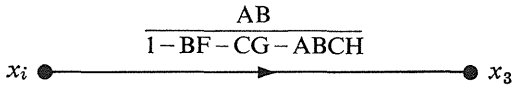
correct statements of the signal relationships in the simple voltage divider on page 99.

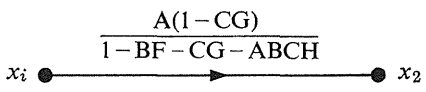
For graphs 1 and 2, above, solve for the graph transmittances A and G_{in} expressing the output voltage in terms of the input voltage, and the input current in terms of the input voltage: (Answer)

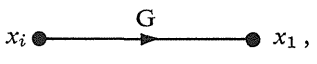


ANSWER TO QUESTION 2.26 Now you must appreciate the ease with which Mason's method enables you to express the relations between signals in a flow-graph. There are evidently three loops in this graph, ABCH, BF, and CG. By applying the method discussed in the previous sections, we find,



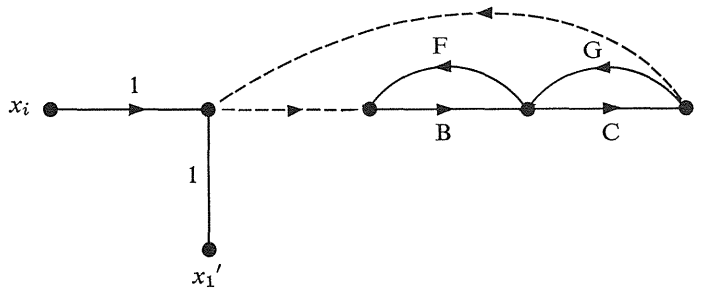






where $G=\dfrac{1(1-BF-CG)}{1-BF-CG-ABCH}$

Note that in calculating the graph transmittance from x_i to any dependent node, we must remember to include in the numerator the cofactor of each path (i.e., the graph determinant associated with the partial graph obtained by erasing all branches of the graph that touch the path). Thus, for the last case considered,

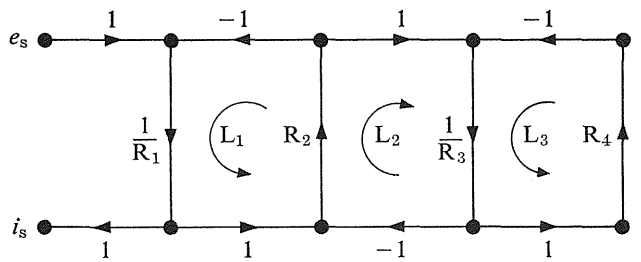


Hence,

$$\Delta_1 = 1 - BF - CG,$$
$$P_1 = 1.$$

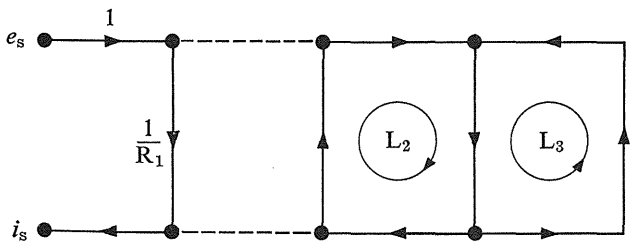
ANSWER TO QUESTION 2.27

1. There are three loops in this graph:



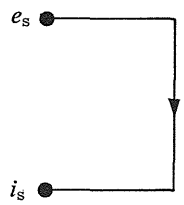
$$L_1 = -R_2/R_1$$
$$L_2 = -R_2/R_3$$
$$L_3 = -R_4/R_3$$

2. There is only one open path, from e_s to i_s. Therefore, the numerator of G is



$$\frac{1}{R_1}(1 - L_2 - L_3).$$

The graph determinant is $1 - L_1 - L_2 - L_3 + L_1L_3$.



$$G = \frac{1}{R_1} \cdot \frac{1-L_2-L_3}{1-L_1-L_2-L_3+L_1L_3}$$

3. If

$$R_1 = 2 \text{ ohms}, \qquad L_1 = -\frac{2 \cancel{\text{ohms}}}{2 \cancel{\text{ohms}}} = -1$$

$$R_2 = 2 \text{ ohms}, \qquad L_2 = -\frac{2 \cancel{\text{ohms}}}{1 \cancel{\text{ohm}}} = -2$$

$$R_3 = 1 \text{ ohm}, \qquad L_3 = -\frac{3 \cancel{\text{ohms}}}{1 \cancel{\text{ohm}}} = -3$$

$$R_4 = 3 \text{ ohms},$$

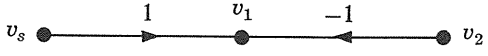
The loop transmittances are *pure numbers* (i.e., dimensionless ratios).

4. Since the loop transmittances are dimensionless, they would not be altered if *all* resistance values were multiplied by the *same* constant. The path transmittance $1/R_1$ would clearly be multiplied by $1/377$ if R_1 were multiplied by 377.

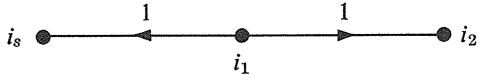
5. $G = \frac{1}{2} \cdot \frac{1 + 2 + 3}{1 + 1 + 2 + 3 + 3} = \mathbf{0.3}$ (unit is *ohms*$^{-1}$ or *mhos*).

ANSWER TO QUESTION 2.28

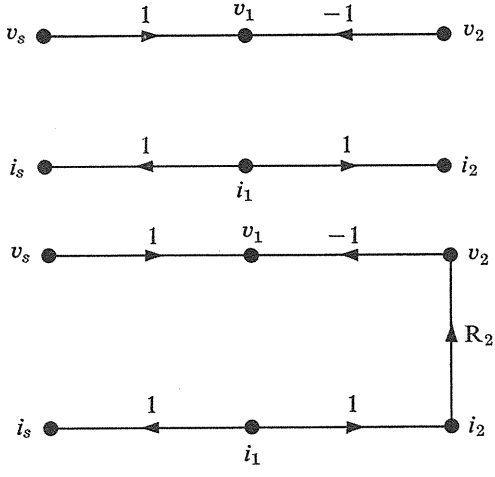
1. Since $v_s - v_2 = v_1$, we have



2. Since $i_1 = i_s$ and $i_1 = i_2$, we have



3. Now the graph looks like



Evidently we may place a scalor branch from i_2 to v_2 since Ohm's law states that $i_2R_2 = v_2$.

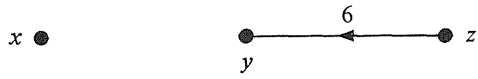
This yields the graph shown.

Next, we come to the relationship between v_1 and i_1. Here an important point must be emphasized. In the graph immediately above, we have already expressed v_1 in terms of v_s and v_2. If we were to draw any additional branches pointing *into* the node v_1, *the signal at that node would no longer be given by* $v_s - v_2$. (It would include other terms, due to the signal flow over the added branches.) But, Kirchhoff's law still requires that $v_1 = v_s - v_2$. Hence, *the graph would be in error!* Once the signal at a node has been defined in terms of other signals, we may add as many *outgoing* branches from that node as we please (this in no way changes the node signal), but *we must not add additional incoming branches.*

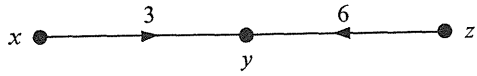
The point of the preceding discussion may be illustrated by a simple example. Consider three signals, x, y, and z, which are related by the equations

$$6z = y \qquad \text{and} \qquad 3x = y.$$

The first relation may be expressed in flow-graph form by



but it would be *incorrect* to *then* express the relation between x and y as



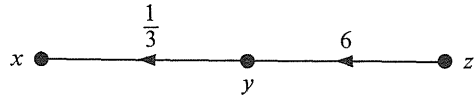
for this graph is equivalent to the equation

$$3x + 6z = y,$$

which, clearly, is *incorrect.*

The correct procedure is as follows. Having already expressed y in terms of z, we must express x in terms of y:

$$x = \tfrac{1}{3}y, \qquad \text{or}$$

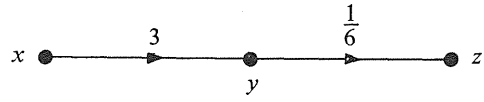


This graph is in accord with the original relations since it states that

$$6z = y,$$

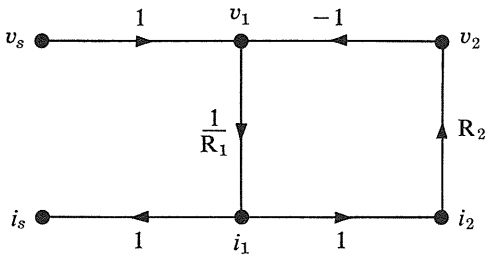
$$\tfrac{1}{3}y = x.$$

Alternatively, we could first express y in terms of x, then x in terms of y:



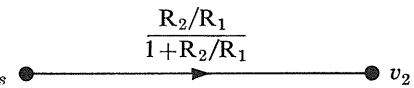
From this example, it should be evident that one cannot add branches indiscriminately to a flow-graph—every branch entering a node may affect the value of the signal at that node. (In contrast, any number of *outgoing* branches may be added to a node without changing the node signal in any way.)

Returning then to the graph for the voltage-divider circuit, we may incorporate an *outgoing* branch from the v_1 node without destroying the relationship $v_1 = v_s - v_2$. Furthermore, we may rearrange the relation $i_1 R_1 = v_1$ so that it reads $v_1/R_1 = i_1$, thus defining i_1 in terms of v_1. This yields the accompanying graph.

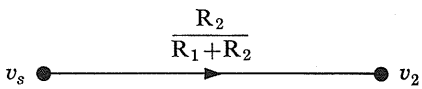


v_s is the source node, and all other signals are dependent.

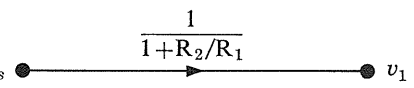
4. By Mason's method, we may solve for v_2 and v_1



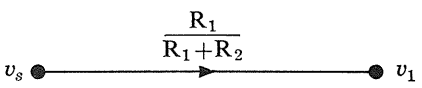
or



and



or



This result is usually described by saying that the *voltage across two resistors in series divides between them in proportion to their respective resistance values.*

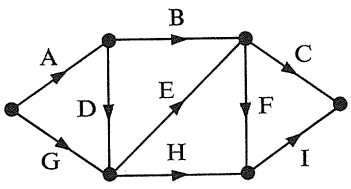
5. Here, application of Mason's method gives

$$G = \frac{1}{R_1}\left[\frac{1}{1+R_2/R_1}\right] = \frac{1}{R_1+R_2}$$

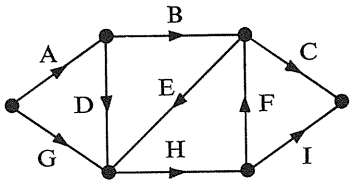
Thus, the two resistors in series draw the same current from the battery as would a single resistor whose resistance is $R_1 + R_2$. This is usually described by saying that *two resistors in series* (so that the current through both is necessarily the same) *act like a single resistor whose resistance is the sum of the respective resistances.*

QUESTION 2.31 For *each* of the following graphs:

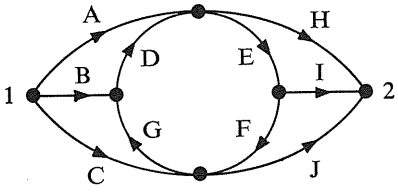
1. Indicate the number of essential nodes (i.e., the minimum number of nodes that must be split or "killed" to interrupt *all* feedback loops).
2. Determine the number of feedback loops that are present in the graph.
3. Give the number of different *nontouching pairs* of loops; the number of *nontouching triplets* of loops; etc.
4. Find the graph transmittance from 1 to 2 for graphs 3, 4, 5, and 6. (Answer)



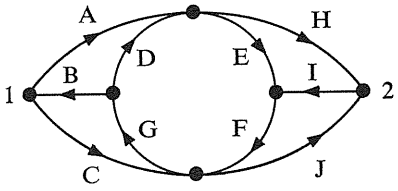
(1)



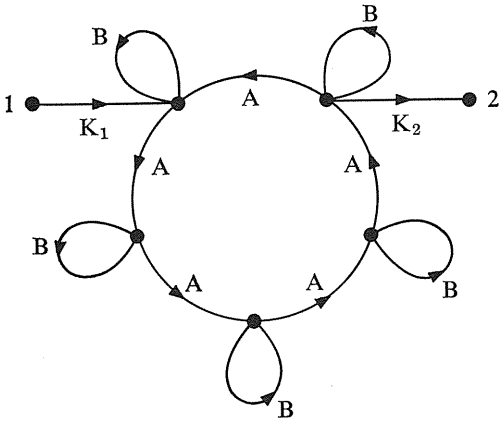
(2)



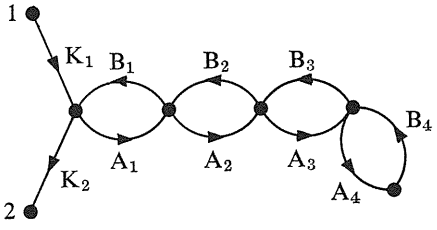
(3)



(4)



(5)



(6)

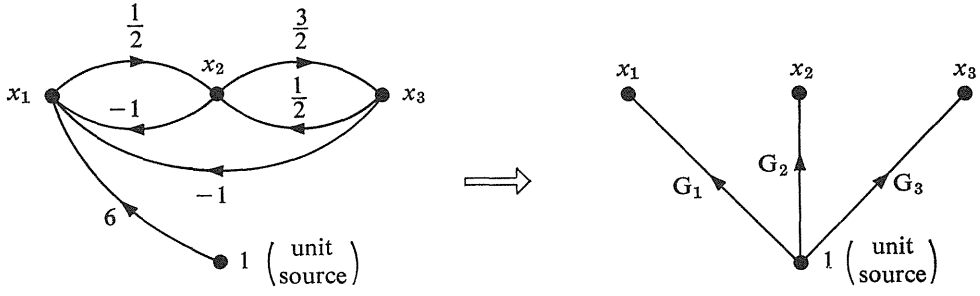
ANSWER TO QUESTION 2.29 This problem illustrates that the flowgraph is an alternative representation for a set of simultaneous equations, and that Mason's method now provides another way of solving such equations.

As the first step, we should rearrange the equations so that *each of the quantities is described explicitly by one and only one equation*:

$$-x_2 - x_3 + 6 = x_1,$$

$$\tfrac{1}{2}x_1 + \tfrac{1}{2}x_3 = x_2,$$

$$\tfrac{3}{2}x_2 = x_3.$$

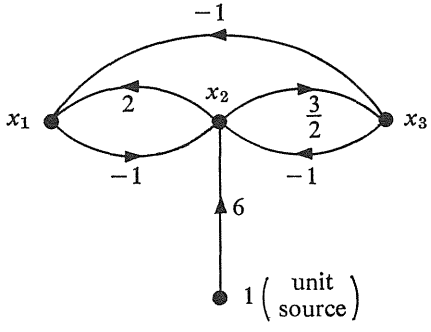


where by Mason's formula

$$G_1 = \frac{6(1 - 3/4)}{1 - (-1/2 + 3/4 - 3/4)} = 1,$$

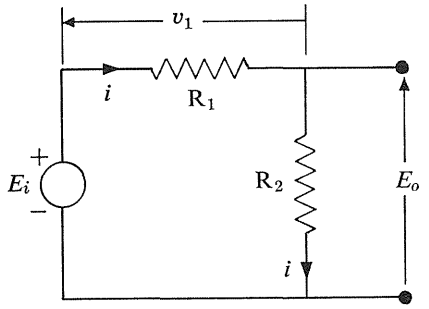
$$G_2 = \frac{(6)(1/2)}{3/2} = 2, \qquad G_3 = \frac{(6)(1/2)(3/2)}{3/2} = 3.$$

You may have rearranged the equations differently and obtained a different flow-graph. For instance, the first equation could have been solved for x_2, the second for x_1, and the third for x_3. This would have given



The solution for x_1, x_2, and x_3 will yield, however, the same values as above.

ANSWER TO QUESTION 2.30 Each of the four flow-graphs shown *is* a *correct* statement of the relationships between the signals of the circuit. To relate to previous work, note that $v_2 = E_o$.



(a)

Kirchhoff's voltage relation

1. $E_i = v_1 + E_o$.

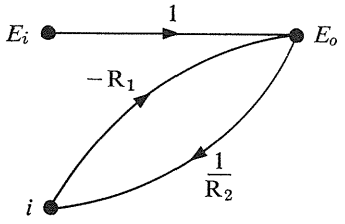
Kirchhoff's current relations

2. All currents are equal.

Current–Voltage relations

3. $v_1 = R_1 i$
4. $E_o = R_2 i$

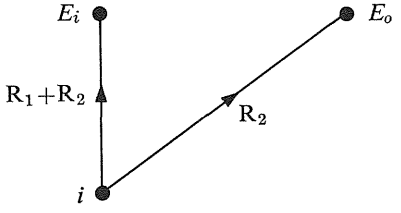
(1)



(b)

	OK by
$E_o = E_i - iR_1$	(1), (3),
$i = E_o R_2^{-1}$,	(4)

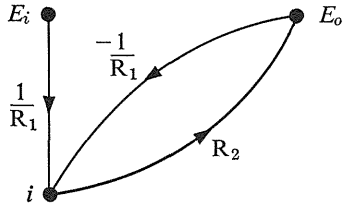
(3)



(c)

	OK by
$E_i = i(R_1 + R_2)$	(1), (3), (4)
$E_o = iR_2$	(4)

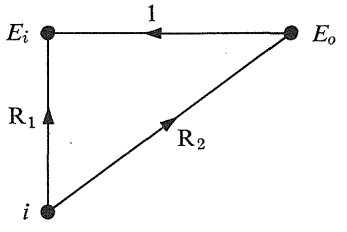
(2)



(d)

	OK by
$E_o = iR_2$	(4)
$i = (E_i - E_o)R_1^{-1}$	(1), (3)

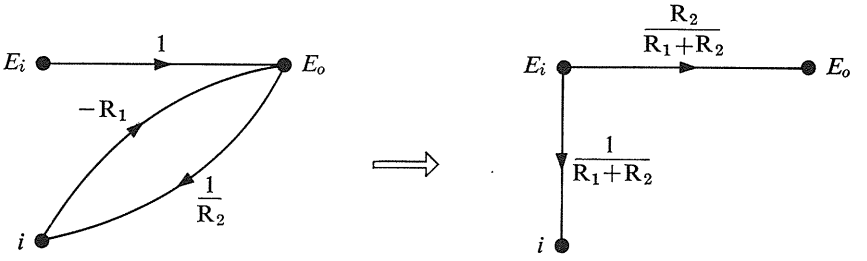
(4)



(e)

	OK by
$E_i = R_1 i + E_o$	(1)
$E_o = R_2 i$	(4)

For graph (1) by direct evaluation of the graph transmittances from the source E_{in} to each of the dependent nodes

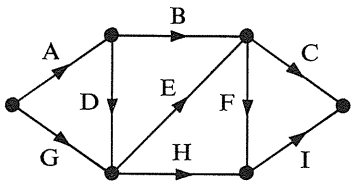


(f)

Reduction of graph (2) also leads to this same result.

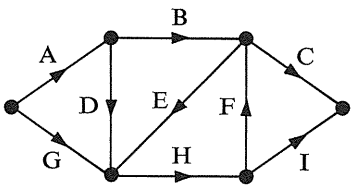
ANSWER TO QUESTION 2.31 Let the minimum number of nodes that must be split to open all feedback loops be called the *graph index*. Let the number of different loops be l_1; the number of different nontouching pairs of loops be l_2; the number of different nontouching triplets of loops be l_3; etc. then,

(1)



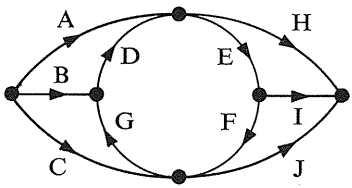
Index: 0
$l_1 = 0, \quad l_2 = 0, \quad l_3 = 0$

(2)



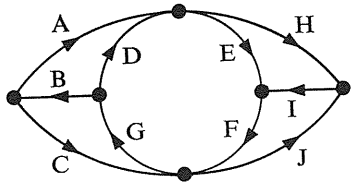
Index: 1
$l_1 = 1, \quad l_2 = 0, \quad l_3 = 0$

(3)



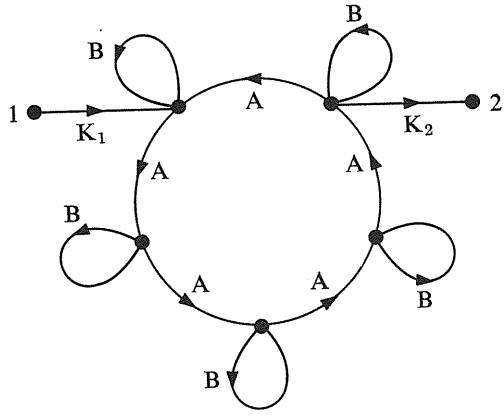
Index: 1
$l_1 = 1, \quad l_2 = 0, \quad l_3 = 0$

(4)



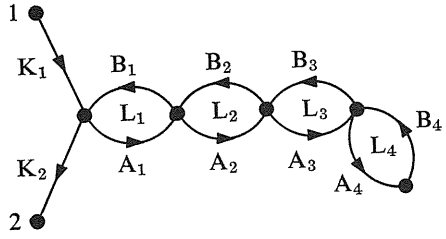
Index: 1
$l_1 = 6, \quad l_2 = 0, \quad l_3 = 0$

(5)



Index: 5
$l_1 = 6$
$l_2 = 10$ (pairs of non-touching loops)
$l_3 = 10$ (triplets of non-touching loops)
$l_4 = 5$ (quadruplets of non-touching loops)
$l_5 = 1$ (quintuplets of non-touching loops)
$l_6 = 0$

(6)



Index: 2
$l_1 = 4$
$l_2 = 3$
$l_3 = 0$

The transmittances from 1 to 2 are:

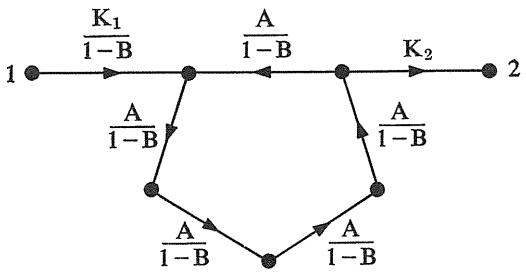
Graph (3)

$$G_{12} = \frac{[A + (B + CG)D][H + EI] + [A + BD]EFJ + CJ}{1 - DEFG}.$$

Graph (4)

$$G_{12} = \frac{[A + CGD][H] + AEFJ + CJ}{1 - BCG - BAEFG - IFJ - IFGDH - BAHIFG - DEFG}.$$

Graph (5) Here, it simplifies matters to eliminate self-loops.



$$G_{12} = \frac{K_1K_2}{1 - B} \cdot \frac{\left(\frac{A}{1 - B}\right)^4}{1 - \left(\frac{A}{1 - B}\right)^5},$$

$$G_{12} = K_1K_2 \frac{A^4}{(1 - B)^5 - A^5}.$$

Graph (6)

$$G_{12} = \frac{K_1K_2(1 - L_2 - L_3 - L_4 + L_2L_4)}{1 - L_1 - L_2 - L_3 - L_4 + L_1L_3 + L_1L_4 + L_2L_4}.$$

Path and Loop Inversion

Discussion of Questions 2.28, 2.29, and 2.31 shows that a flow-graph may be transformed into many different forms. A given set of equations may also be rearranged and written in many different ways, and these rearrangements correspond exactly with alternate forms of a flow-graph. This suggests that we should be able to rearrange a flow-graph from one form to a number of other equivalent forms if we establish the proper rules.

An important transformation is the *inversion of a path or loop.* This transformation is useful for interchanging the "cause" and "effect" roles of two nodes in a graph. We have already seen an example of this for a simple resistance.

In Figure 2.56, i is taken as the "cause" of e. Here, i is a *source* node and e is a dependent node. To express the dependency of i on e, we must *reverse* the direction of the branch *arrow* and replace the transmittance by its *reciprocal.* The validity of this transformation is evident by comparing the equations, $e = Ri$ and $i = R^{-1}e$.

FIGURE 2.56

With more branches the process is a little more complicated than in this example. Consider Figure 2.57. In this graph, x_1 and x_3 are shown as sources which may take on any desired values. Once the values of x_1 and x_3 have been specified, however, the

values of all other node signals are determined by the graph, in accordance with the equations:

$$x_2 = x_1A_1 + x_4B,$$
$$x_4 = x_2C + x_3D,$$
$$x_5 = x_4E.$$

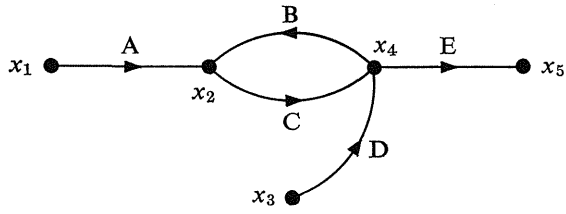


FIGURE 2.57

Note that *one* equation is explicitly written for each *dependent* signal encountered along a path leading from x_1 to x_5.

Suppose that we wish to specify independently the value of x_3 and also of x_5. The value of x_1 can no longer be chosen independently. Instead, x_1 must take on that value which would yield the *desired value* of x_5. Thus, x_1 now depends on x_5. The effect is to convert x_5 into a *source* node and x_1 into a *dependent* node. The new dependent nodes are x_1, x_2, and x_4.

A corresponding rearrangement of the equations yields

$$x_1 = [x_2 - x_4B]A^{-1},$$
$$x_2 = [x_4 - x_3D]C^{-1},$$
$$x_4 = x_5E^{-1},$$

and the corresponding graph, illustrated by Figure 2.58, shows where an extra node has been introduced to represent each of the bracketed signals, $x_2 - x_4B$ and $x_4 - x_3D$, in the equations. In the first bracket, the term $-x_4B$ is identical (except for the *minus* sign,

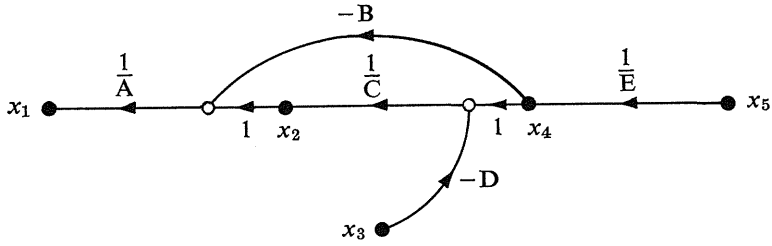


FIGURE 2.58

which arises when it is transposed to the other side of the equation) to the term x_4B in the original equation. Correspondingly, the branch B of the original graph is shifted to the

parenthetical node representing these bracketed signals, and the algebraic sign of its transmittance is changed. Likewise, when the equation $x_4 = x_2\mathrm{C} + x_3\mathrm{D}$ is solved for x_2, the term $x_3\mathrm{D}$ is transposed to the other side of the equation, creating the parenthetical quantity $(x_4 - x_3\mathrm{D})$. Correspondingly, the original branch D is moved to the parenthetical node and the algebraic sign of its transmittance is changed. Evidently, the transmittances of the branches A, C, and E along the original path are replaced by their inverses when the path is inverted.

The process just described in connection with the particular graph may be applied generally to invert *any path* which starts at some *source* and terminates at a dependent node. It may be applied even when the branch operators are nonlinear, provided that A^{-1} denotes the inverse nonlinear operator. After the inversion process has been carried through, the dependent node will have become a source, and vice versa.

When the operators are linear, it is possible to absorb the "parenthetical" nodes that were introduced in the inversion process by eliminating the parentheses from the equations:

$$x_1 = \frac{1}{\mathrm{A}} x_2 - \frac{\mathrm{B}}{\mathrm{A}} x_4,$$

$$x_2 = \frac{1}{\mathrm{C}} x_4 - \frac{\mathrm{D}}{\mathrm{C}} x_3,$$

$$x_4 = \frac{1}{\mathrm{E}} x_5.$$

This yields Figure 2.59.

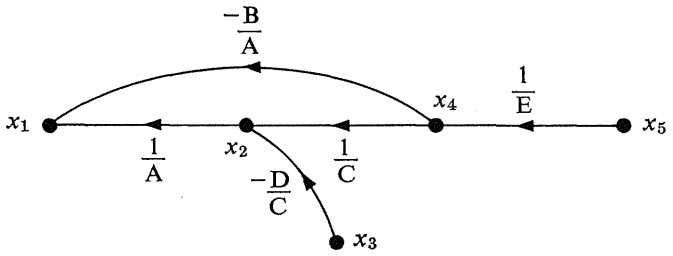
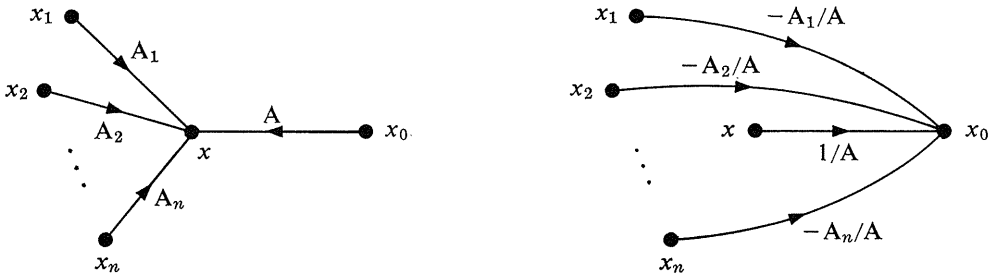


FIGURE 2.59

By generalizing the process illustrated in this example, we may now state the general rule for inverting an open path: begin with the branch leaving the original source, reverse the direction of the first branch, and replace the transmittance by its reciprocal. Carry with the output end of the branch being inverted the output ends of all other branches that originally terminated in this same node, and multiply the transmittance of each of these branches by the negative of the new transmittance of the inverted branch. Repeat this process with the second branch of the open path, and continue in this manner until all branches of the open path have been inverted. This operation, difficult to describe

clearly in words, is actually very simple to perform. Figure 2.60 illustrates this process by inverting a single branch A.



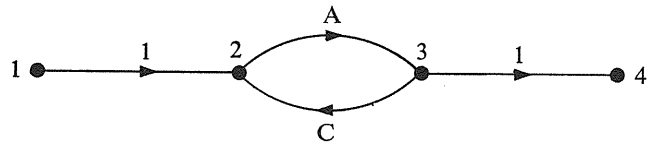
$$x = x_0A + x_1A_1 + x_2A_2 + \cdots + x_nA_n \qquad x_0 = [x - x_1A_1 - x_2A_2 - \cdots - x_nA_n]\frac{1}{A}$$

FIGURE 2.60

In addition to inverting open paths, it is possible to invert a loop. To do this correctly, it is helpful *first to split one of the nodes in the loop*, thus temporarily transforming the loop into an open path between a source and a sink. After this path has been inverted, the split node is rejoined.

We should carefully define what we mean by *node splitting*. The process of splitting a node divides that node into a new *source* and a new *sink*, with all *incoming* branches remaining attached to the new *sink*, and all *outgoing* branches remaining attached to the new *source*. The result of splitting a node is to block all signal flow through the node (i.e., the node is "killed"). Since this process produces two nodes for each one that is split, we must agree on a notation for designating the new nodes. Following Mason (see Reference 13), we shall retain the original signal symbol and the original node num-

QUESTION 2.32 You are to invert a path through this graph so as to make node 4 a source and node 1 a sink.

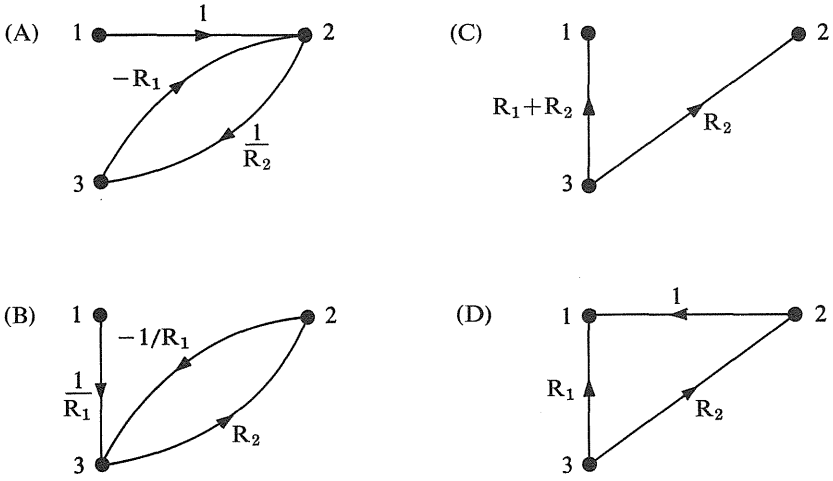


In this graph there is only one open path from 1 to 4. This path involves three branches. Which one of these branches *must* be inverted first? Which branch is inverted next? Which branch is inverted last? (Answer)

QUESTION 2.33 In Question 2.30, four *different* flow-graphs were given, each of which correctly represented the relationships between the signals E_i, E_o, and i in a resistance voltage-divider circuit. Since each of these flow-graphs describes the *same*

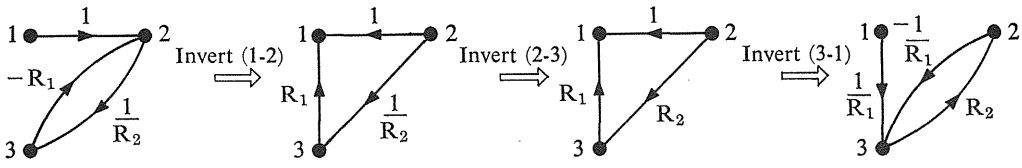
physical situation, each must convert to the other by applying any of the transformations that relate equivalent flow-graphs. Path inversion is one of the most important transformations.

The four flow-graphs are repeated here. Their nodes are numbered to permit easy description of any path through the graph.



In this problem, we are concerned with the conversion of each of these graphs into the others when possible by means of path inversion transformations. For instance, graph A may be converted to graph E as follows:

Graph A



Graph B

1. Which of the graphs other than B can be obtained from graph A by use of the path-inversion transformation?
2. Which of the graphs can be obtained from graph B by use of the path-inversion transformation?
3. Which of the graphs can be obtained from graph C?
4. Which of the graphs can be obtained from graph D?
5. Are there any graphs that cannot be obtained in the previous questions? If so, what equivalence transformation is needed so that every graph can be expressed in terms of every other?

ber for the *new source*, whereas the corresponding quantities at the *new sink* will be designated by primes. Thus, if we split node 2 in Figure 2.61 at the left, the resulting graph appears as at the right. Once a node is split, the graph may be transformed in various

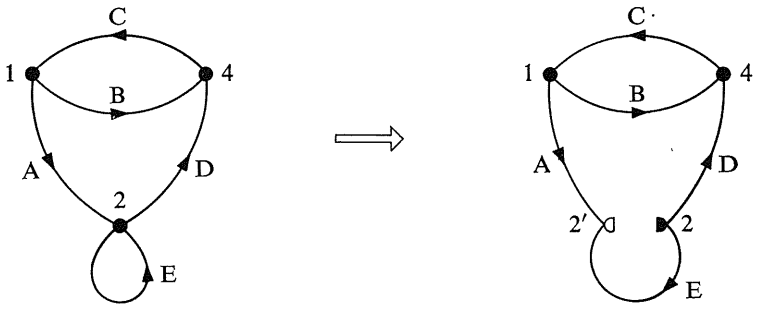


FIGURE 2.61

ways. For instance, the total transmittance from node 2 to node 2′ may be found by absorbing nodes 1 and 4. This transmittance reveals how much signal is returned to node 2 per unit of signal at that node. The notion of loop transmittance at a node, obtained in this way, will be used later in the proof of Mason's theorem.

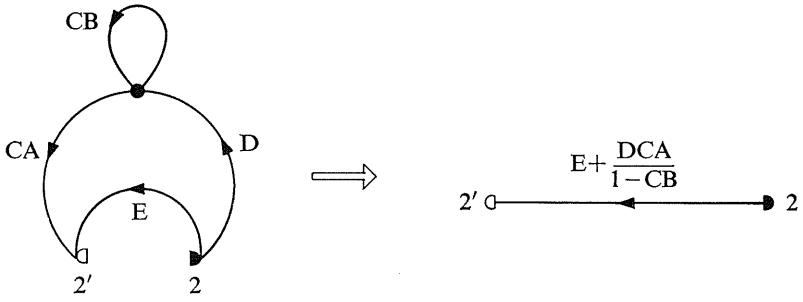
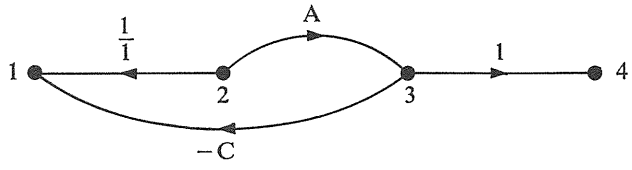
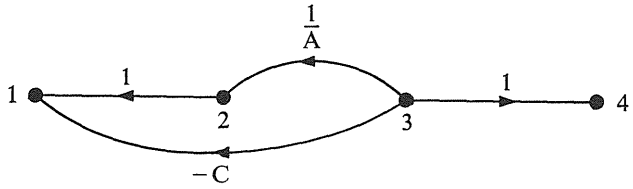


FIGURE 2.62

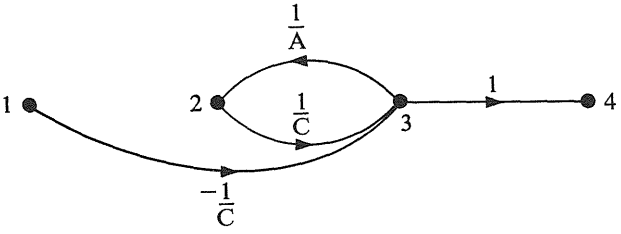
ANSWER TO QUESTION 2.32 Remember that path inversion makes a *source* node into a *dependent* node, and the *dependent* node into a *source* node. Because of this, you *cannot* invert a branch that connects two *dependent* nodes; a branch can be inverted only if it begins at a *source node*. Thus, of the three branches forming the path from 1 to 4, only the branch from 1 to 2 may be inverted. This yields



But now, node 2 has become a *source*. Any branch leaving node 2 could be inverted. Inversion of the branch from 2 to 1 would return the graph to its original condition. Inversion of the branch from 2 to 3 yields



But now, node 3 is a *source*. Any branch leaving node 3 may next be inverted. For instance, the branch from 3 to 1 could be inverted, making node 1 into a source again:

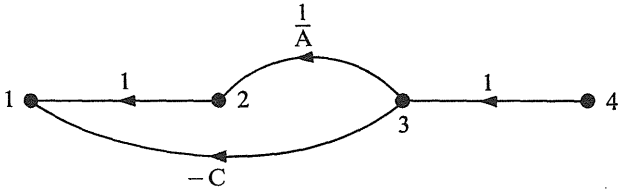


This graph is *equivalent* to the original graph since the signals at corresponding nodes are identical. Thus, G_{14} is

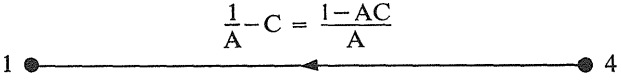
$$\frac{-1/C}{1-(1/A)(1/C)} = \frac{A}{1-AC},$$

where the expression at the right is easily seen to be G_{14} for the original graph.

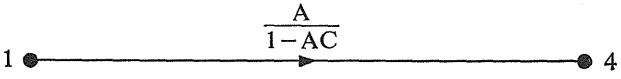
To make node 4 a source, we invert the other branch leaving the source node 3, to obtain



The graph transmittance from 4 to 1 is evidently

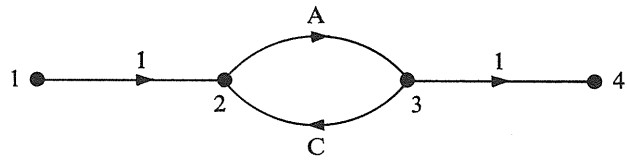


whereas in the original graph the transmittance from 1 to 4 was



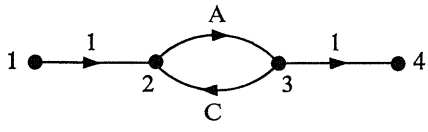
These transmittances are clearly inverses of each other.

QUESTION 2.34 For the graph shown here, what is the graph obtained by splitting node 3? (Answer)

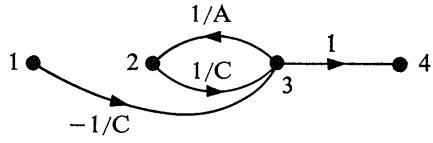


QUESTION 2.35 By splitting a node, one creates a source node and a sink node. What is the graph obtained by inverting the path from source node 3 to node 3′? (Answer)

By reconnecting the two parts of the split node in the graph obtained in Question 2.35, we obtain a graph that is equivalent to the original graph:



Original graph



Graph obtained by inverting loop

FIGURE 2.63

Loop inversion is important because it sometimes reduces markedly the number of loops in the graph and may make it much easier to solve. For instance, inversion of the center loop of graph 5 of Question 2.31 (previously considered) yields the graph at the left of Figure 2.64. Even though these graphs look quite different, the source-to-sink transmittances are identical. Inversion of a loop leaves the source and sink nodes unchanged, and the two graphs are equivalent because the signals are the same for both graphs.

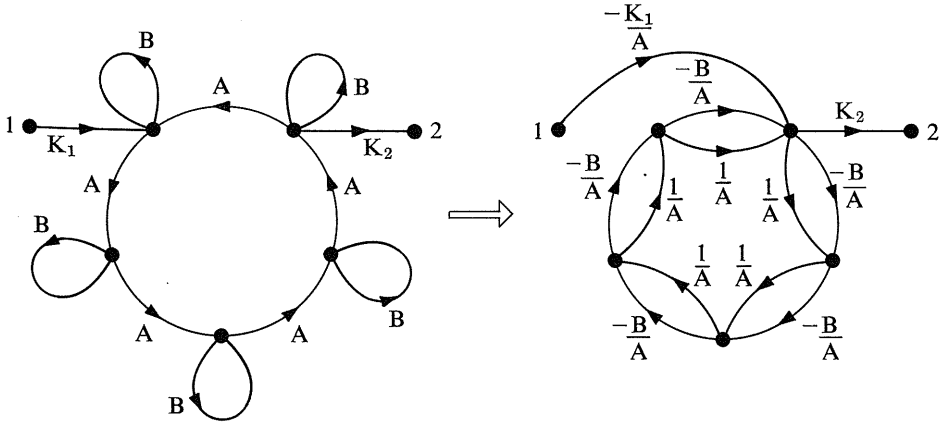
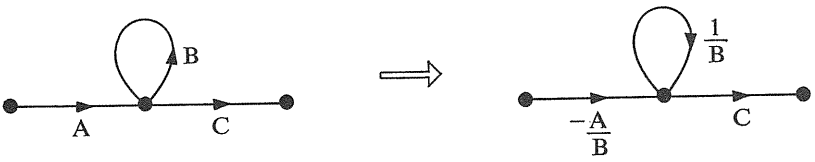


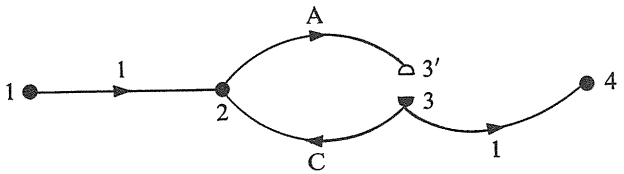
FIGURE 2.64

Path inversion is sometimes applied to a *self*-loop. The necessity for inverting a loop often arises in formulating a model for solution on an analog computer. The advantages of inverting a self-loop are illustrated by Question 2.36.

QUESTION 2.36 With reference to the simple graph shown at the left, can the graph at the right be obtained by inverting the self-loop? (Answer)

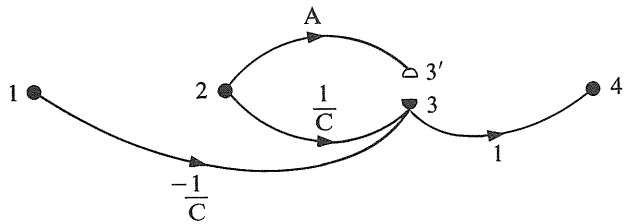


ANSWER TO QUESTION 2.34 To split a node we move all incoming branches to one part of the split node and leave the outgoing branches with the other part:

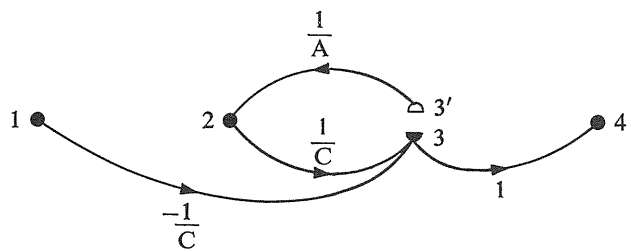


The node 3 thus becomes a source, and node 3′ a sink.

ANSWER TO QUESTION 2.35 The path is (3, 2, 3′). Inversion of branch (3, 2) gives



Inversion of branch (2, 3′) gives

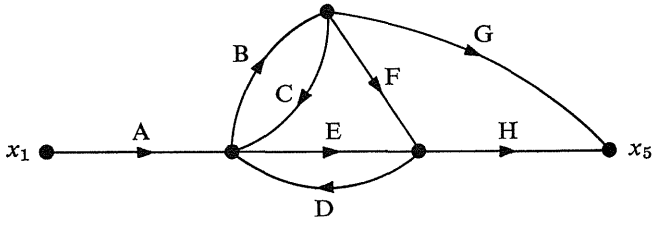


The various flow-graphs obtained by inverting different paths correspond to the various arrangements of a set of algebraic equations when expressed in cause–effect form, with one equation explicitly representing each variable. By this inversion process, all of the cause-and-effect relationships may be obtained without further recourse to the equations once a problem is in flow-graph form.

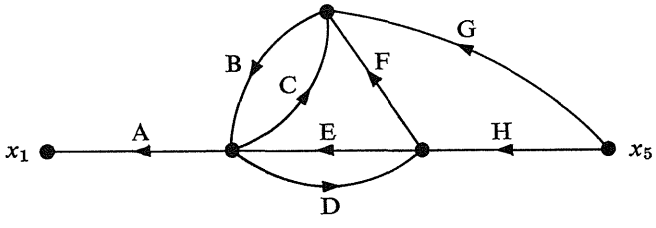
The Converse of a Flow-Graph

Our study of flow-graph transformations, equivalences, and reductions is now nearly completed. One topic remains—the converse of a flow-graph.

We shall define the *converse* of a given flow-graph as that graph which is *identical* to the given graph *except that the directions of the arrows on all the branches have been reversed.* Specifically, if B_{jk} is the transmittance from node j to node k in the original graph, then the transmittance B'_{kj} from node k to node j in the converse graph is numerically equal to B_{jk}. This is illustrated by the graphs in Figure 2.65.



Original graph



Converse graph

FIGURE 2.65 *The converse graph is formed by reversing all the arrows.*

QUESTION 2.37 Consider now the source-to-sink transmittance G_{15} of the original graph in Figure 2.65, and compare it to the source-to-sink transmittance G'_{51} of the converse graph. What can you say about these two transmittances? (Answer)

Appendix A: Proof of Mason's Loop-Expansion Theorem

Evidently Mason's loop-expansion theorem works, and works well. Why it works we have not yet shown. In this Appendix, we outline a proof of this theorem as given in Reference 13 in Appendix B. The proof consists of the following four parts:

1. Definition of the graph determinant in terms of partial loop transmittances at each node of the graph.
2. Demonstration that the graph determinant is independent of the order in which the nodes are numbered.
3. Demonstration that when a new node is added to the graph, the determinant of the enlarged graph may be expressed in terms of the loops passing through the new node and the determinant of the old graph.
4. Obtaining the graph transmittance formula by splitting the new node considered in part 3, and identifying the two parts as the source and sink nodes associated with the graph transmittance.

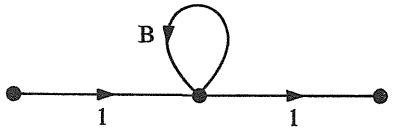
1. Definition of graph determinant in terms of partial loop transmittances—In the previous section, we discussed *node splitting* briefly. Node splitting is an artifice to help keep the incoming and outgoing signals separated at a node. Splitting a node is a way of "killing" a node in the sense that incoming signals then produce no outgoing signals. For instance, splitting node 2 in the graph at the left leads to the graph in the center in

ANSWER TO QUESTION 2.36 Yes, it is obtainable.

The important point here is that if B is greater than unity, then 1/B will be less than unity.

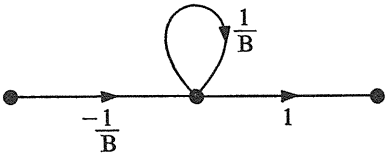
One important application of loop inversion occurs with analog computers. It is sometimes desirable to avoid loops having transmittances that are much larger than unity, for this condition may produce undesirable spurious oscillations in the computer. By inverting such a loop, we obtain a new graph with loop transmission much smaller than unity. This may be realized more readily on a computer.

The difficulty here is similar to that already encountered in the power-series interpretation of the effect of a self-loop (see p. 63).



$$G = 1 + B + B^2 + B^3 + \cdots$$
$$= \frac{1}{1 - B}.$$

The infinite geometric series converges only if $B < 1$, whereas if $B = 1$, the output signal is infinite for any nonzero input signal. But by inverting the self-loop,



$$G = \frac{-1}{B}\left[1 + \frac{1}{B} + \frac{1}{B^2} + \frac{1}{B^3} + \cdots\right]$$
$$= -\frac{1}{B\left(1 - \frac{1}{B}\right)}, \quad \text{which is equivalent to}$$
$$= \frac{1}{1 - B},$$

one obtains a convergent series for all $B > 1$.

Figure 2.66. The graph transmittance from the source to sink nodes thus formed is defined as the *loop transmittance at a node*, shown in Figure 2.66.

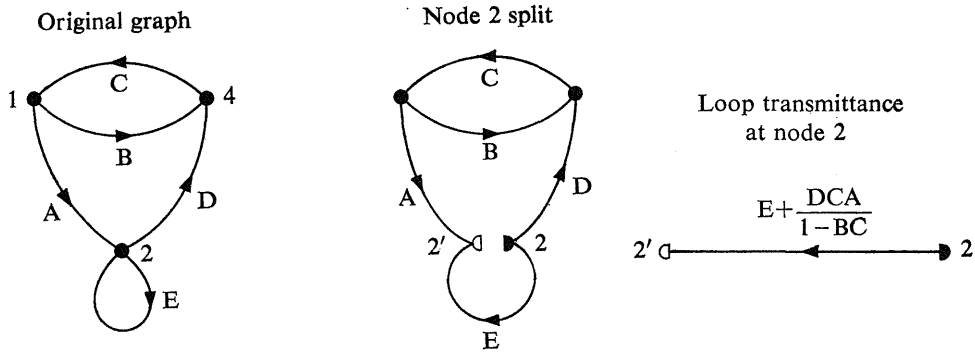
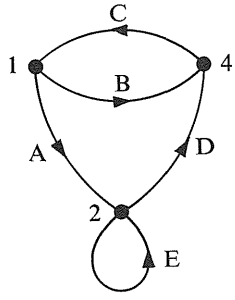


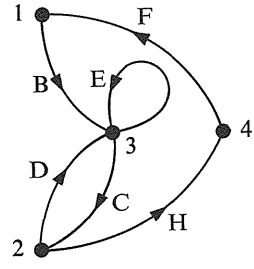
FIGURE 2.66 *By splitting node 2, the loop transmittance may be found at node 2.*

Since splitting a node blocks all signal flow through it, it is evident that if we split *all* the nodes in the graph the graph will be completely dead. There certainly would be no loops remaining. However, it generally is *not* necessary to split *all* the nodes to open all the loops in the graph—a smaller subset of nodes may often be selected which when split will open all loops originally present. If there is *no* subset containing fewer nodes, the subset is said to form a *set of essential nodes.* That is, a set of essential nodes is the smallest set of nodes which when split will open all loops in the original graph. There may be several different sets of essential nodes for a given graph, but the number of essential nodes, called the *index* of the graph, is the same for each set.

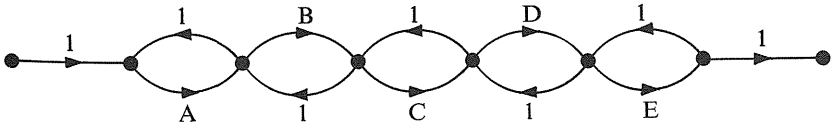
QUESTION 2.38 For each of the three accompanying graphs, what is the smallest number of nodes that would have to be split to open all loops of the graph? In each case, list all possible sets of essential nodes, and give the index of each graph. (Answer)



(1)



(2)



(3)

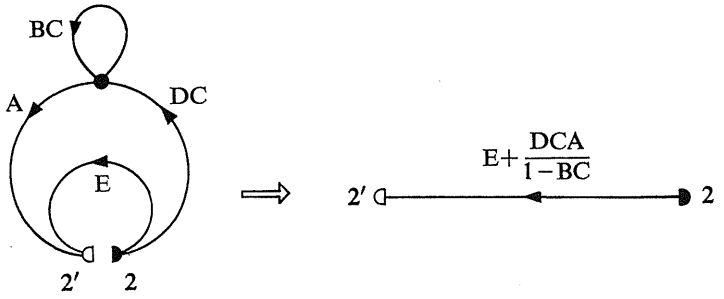


FIGURE 2.67

In the example considered in Figure 2.66, the loop transmittance at node 2 represents the signal that returns to that node per unit of signal sent out by that node. When the split node is rejoined, the single branch becomes a loop. We thus define a *loop transmittance* T_j for any node j as

T_j = transmittance between the new source–sink pair created by splitting node j.

Thus, the loop transmittance at node 2 is as shown at the left in Figure 2.68. The *graph*

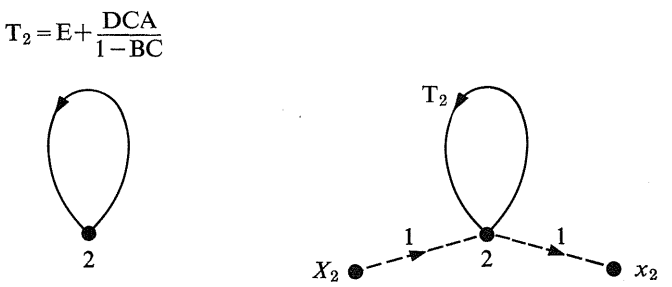


FIGURE 2.68

transmittance G_{22} is found, by attaching an external source and sink as shown at the right in Figure 2.68, to be

$$G_{22} = \frac{1}{1 - T_2} = \frac{1 - BC}{1 - E - BC - DCA + EBC}.$$

ANSWER TO QUESTION 2.37 It is true that $G_{15} = G'_{51}$. This rather remarkable result would hardly be expected from a comparison of the two sets of simultaneous equations that the two graphs represent. However, once these equations have been put into flow-graph form, the equivalence is easily proven using Mason's method. Changing the directions of the arrows on *all* of the branches leaves the paths from a specified source to a specified sink unchanged. Furthermore, all the loops remain unchanged. Clearly, the path transmittances and loop transmittances of the original graph will be identical to the corresponding path and loop transmittances of its converse. Hence, the graph transmittances are identical. This possibility of replacing one system by *another* system which has the *same* solution is often exploited in linear programming (a technique for optimization used in Operations Research).

Again, the same expression for the graph determinant emerges in the denominator of the graph transmittance. We might expect that there would be some relation between the loop transmittances and the graph determinant. This is indeed so. In fact, by definition, the *determinant*, Δ, *of a flow-graph* is given by

$$\Delta = (1 - T_1')(1 - T_2'), \ldots, (1 - T_n'), \tag{2.7}$$

where T_k' is the value of T_k with *the higher-numbered nodes* $k + 1, k + 2, \ldots, n$ *split, so as to block all signal transmission through them.*

Since only part of the graph is active, the primed T_k may be called the *partial loop transmittance* of node k. Evidently, the values of these partial loop transmittances will depend on the order in which we choose to assign numbers to the nodes. However, *the remarkable property of the graph determinant* Δ *is that its value is independent of the order in which the nodes are numbered.*

Now let us see if we can use this definition to evaluate the determinant of the graph in Figure 2.69. First, evaluate the partial loop transmittances of nodes 1, 2, and 4 in

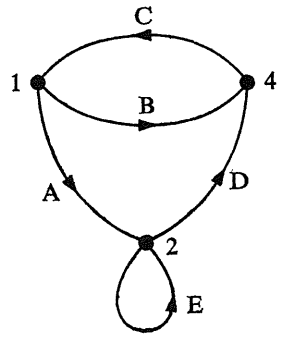


FIGURE 2.69

that order. Remember that in evaluating T_1', nodes 2 and 4 are first split *so there will be no signal transmission through them.* Similarly, in evaluating T_2', only node 4 is split, etc. When these partial loop transmittances are substituted in

$$\Delta = (1 - T_1')(1 - T_2')(1 - T_3')(1 - T_4'),$$

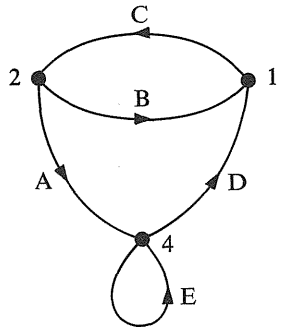
you should obtain

$$\Delta = (1 - 0)(1 - E)(1)\left(1 - BC - \frac{DCA}{1 - E}\right)$$

$$= 1 - E - BC - DCA + EBC,$$

which is the same value as before. Note that since node 3 does not appear in the graph, its loop transmittance T_3' is zero, and hence the difference, $(1 - T_3')$, is unity and has no effect on the value of Δ. Hence, parenthetical factors associated with nodes that do not appear in the graph may be omitted without altering the value of Δ.

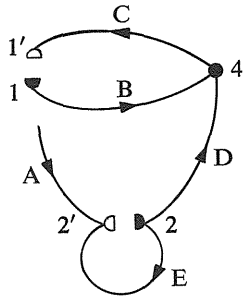
QUESTION 2.39 What are the partial loop transmittances for the same graph, evaluated with the nodes renumbered as follows: (Answer)



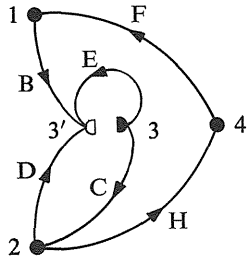
2. Demonstration that the graph determinant is independent of the order in which the nodes are numbered—To see *why* the value of the graph determinant Δ is independent of

ANSWER TO QUESTION 2.38 The smallest number of nodes that would have to be split in order to open all loops of the graph is two nodes in graph 1; one node in graph 2; and three nodes in graph 3, as illustrated.

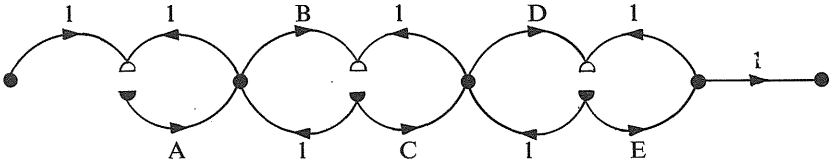
Graph 1: two nodes



Graph 2: one node



Graph 3: three nodes



The choice of the particular set of nodes that must be split to open all loops is not necessarily unique, although the *number* of nodes in this set is unique (Mason calls the number of nodes in this set the *index* of the graph.) For graph 2, all feedback loops may be interrupted by splitting one node in only one way. However, for graphs 1 and 3 there are two different sets of nodes that will have this effect in each graph.

the order in which the nodes are numbered, we need only investigate the effect of interchanging the numbering of *any pair*, let us say nodes k and $k + 1$, of consecutively numbered nodes while leaving the numbering of all other nodes unchanged. If we can show that interchanging the numbering of this *pair* of consecutively numbered nodes leaves the graph determinant unchanged, the general property will have been demonstrated, since *any* node-numbering order can be obtained by making a succession of adjacent interchanges. (If you doubt this, demonstrate to your own satisfaction that the sequence 123 can be transformed into the sequence 321 by making three successive interchanges of adjacent numbers.)

Consider, then, the partial loop transmittance T'_{k+1}. By definition, this transmittance is to be computed from the partial graph containing only nodes 1, 2, 3, . . . , k, and $k + 1$; all other nodes, $k + 2$, $k + 3$, . . . , n being blocked by splitting. If nodes 1 through $k - 1$ are successively absorbed, this partial graph reduces to a form containing only the nodes k and $k + 1$, shown in Figure 2.70. Had we wished to find T'_k, node

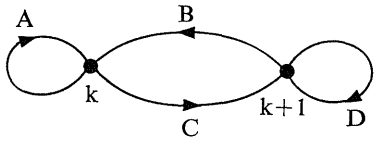


FIGURE 2.70

$k + 1$ would have been split, and clearly

$$T'_k = A.$$

Furthermore, we find that the partial loop transmittance T'_{k+1} may be expressed directly in terms of this reduced graph as $T'_{k+1} = D + BC/(1 - A)$. The product of the two factors appearing in the graph determinant is thus

$$(1 - T'_k)(1 - T'_{k+1}) = (1 - A)(1 - D) - BC.$$

We next show that had we interchanged the numbering of these two nodes, the transmittances of the reduced graph would have remained unchanged. By the same reasoning as before,

$$T'_k = D \qquad \text{and} \qquad T'_{k+1} = A + \frac{CB}{1 - D},$$

yielding for the product of the two terms

$$(1 - T'_k)(1 - T'_{k+1}) = (1 - D)(1 - A) - CB,$$

which is identical to the previous value (provided $BC = CB$).

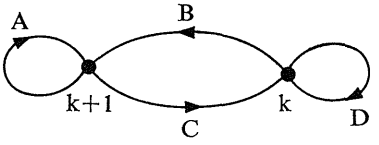


FIGURE 2.71

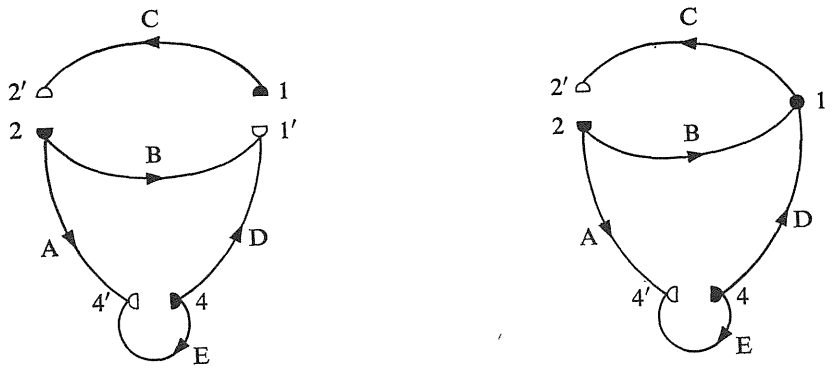
Since adjacent interchanges in the node numbering do not affect the product of the corresponding factors in Δ, and since *any* node numbering can be achieved by a suitable succession of such interchanges, we have proved that the *value of the graph determinant is independent of the order in which the nodes are numbered.*

3. Demonstration that when a new node is added to the graph, the determinant of the enlarged graph may be expressed in terms of the loops passing through the new node and the determinant of the old graph. In this section, we investigate the selection rules that govern the various branch-transmittance products appearing in Δ. (We have already concluded that these terms involve products of *nontouching* loops, but why is this so?)

We have just proved the fundamental property of the graph determinant—that the determinant $\Delta = (1 - T_1)(1 - T_2) \ldots (1 - T_n)$ has the *same value* irrespective of the order in which nodes are numbered. Since this is so, we can always number the n different nodes so that any particular node will be the nth or last node. The computation of the loop transmission for this last node will then involve finding the total transmittance between the two parts of this node when it is split, taking into account the effect of all the other $n - 1$ nodes of the graph.

To illustrate this situation, Mason uses the diagram of Figure 2.72. Now the total transmittance from n to n' will be equal to the sum of the transmittances over the various

ANSWER TO QUESTION 2.39 Initially, all nodes are split. Hence $T_1' = 0$. Then rejoining node 1 and evaluating T_2', we have for the partial loop transmittance at node 2,



$$T_2' = BC.$$

Finally, for node 4 (as before)

$$T_4' = E + \frac{DCA}{1 - BC}.$$

Hence the graph determinant is

$$\Delta = [1 - 0][1 - BC]\left[1 - E - \frac{DCA}{1 - BC}\right]$$

$$= 1 - BC - E - DCA + EBC.$$

paths each of which must pass through one of the branches associated with node n. But note that any path that includes branch A cannot also include branches B or C.

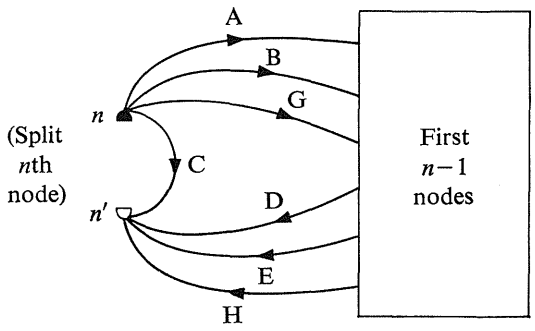


FIGURE 2.72

Likewise, any path that includes D cannot include paths C or E. If we first define two branches to be *confluent* if they both *originate*, or *both terminate* at the same node, then it is evident that the transmittance of *any one* of the paths from n to n' cannot involve the transmittance product of two (or more) branches confluent at node n. The idea here, which is much simpler than it sounds, dictates that the path transmittances could contain transmittance products such as: AE, BD, BH, C, but cannot contain transmittance products such as: AB, A^2, ABG, EHA, etc.

QUESTION 2.40 To make sure you understand the significance of confluent branches, what are six other permissible transmittance products in addition to the four given in the preceding paragraph? (Answer)

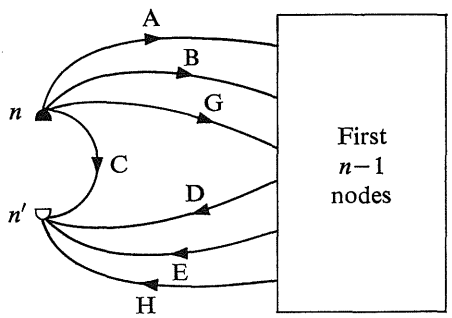


FIGURE 2.73

Let us see what this conclusion means with respect to the graph determinant

$$\Delta = (1 - T_1')(1 - T_2') \ldots (1 - T_{n-1}')(1 - T_n)$$

where we have written T_n for T_n' since there are *only* n nodes in the graph and hence the partial loop transmittance is identical with the loop transmittance at the last node. If we designate the product of the first $n - 1$ factors by Δ', the determinant can be written as

$$\Delta = \Delta'(1 - T_n).$$

Now consider the terms that appear in Δ'. They involve partial loop transmittances which were computed with node n split. Hence, *none* of the branches A, B, G, C, D, E, H that are confluent on node n can appear in any of the terms in Δ. Furthermore, the loop transmittance T_n will contain terms *none* of which contain the product of confluent branches at node n. However, by renumbering the nodes, node n can be made to be *any node* in the graph, so that same conclusion must hold for any node. Thus, we have proved the important property that Δ *consists of the sum of terms none of which contains products on confluent branches.*

The terms appearing in Δ each involve products of transmittances associated with the various closed loops in the graph. When two loops touch at a common node, the transmittance products of the two loop transmittances will necessarily involve two (or more) confluent branches. But we have just proved that such products *cannot* appear in Δ. Hence, *each term of the graph determinant must be a simple product of nontouching loops.* The graph determinant, which systematically incorporates terms arising from all possible products of nontouching loops, is thus expressible as

$$\Delta = [(1 - L_1)(1 - L_2) \ldots (1 - L_n)]^*, \tag{2.8}$$

where "*" denotes that any term containing products of touching loops is to be dropped.

4. Obtaining the graph transmittance formula by splitting the new node considered in part (3) and identifying its two parts as the source and sink nodes associated with the graph transmittance—With this major result in hand, Mason now moves rapidly toward our goal of evaluating the graph transmittance by simple inspection. Suppose that we have calculated the graph determinant Δ associated with the first n nodes. What will be the effect of adding a new node $n + 1$ to the graph?

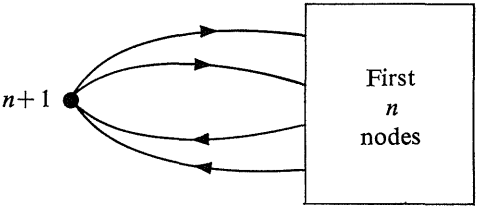


FIGURE 2.74

Clearly, the addition of the new node and its associated branches will produce some number k of new loops L'_v which were not present in the original graph. If we denote the determinant of the complete graph having $n + 1$ nodes by $\Delta^\dagger$, then by applying Mason's loop expansion theorem, we must have

$$\Delta^\dagger = [\Delta(1 - L'_1)(1 - L'_2) \ldots (1 - L'_k)]^* \tag{2.9}$$

But each of the L'_v loops are confluent since they *all* touch one another at node $n + 1$. Hence, in forming the product of the last k factors, only the constant and single-loop

ANSWER TO QUESTION 2.40 AD, AH, BE, GD, GE, and GH are the only other permissible transmittance products.

terms remain. All products of pairs, triplets, etc., of loops must be dropped because the loops are confluent. Hence,

$$\Delta^\dagger = [\Delta(1 - L_1')(1 - L_2') \ldots (1 - L_k')]^* = \left[\Delta\left(1 - \sum_{\nu=1}^{k} L_\nu'\right)\right]^*$$

or

$$\Delta^\dagger = \Delta - \left[\Delta \sum_{\nu=1}^{k} L_\nu'\right]^*,$$

$$\Delta^\dagger = \Delta - \sum_{\nu=1}^{k} L_\nu' \Delta_\nu, \tag{2.10}$$

where Δ_ν is the determinant of the subgraph formed by erasing all branches of the original graph which touch the loop L_ν'. To illustrate, let us consider Figure 2.75.

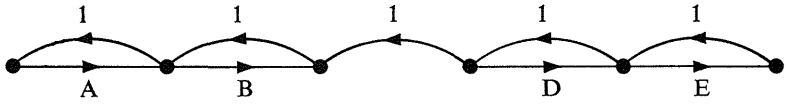


FIGURE 2.75

Here the loop subgraph breaks into parts, so the determinant can be expressed as the product of the determinants of its two parts:

$$\Delta = (1 - A - B)(1 - D - E).$$

Suppose that now we add an additional node, as in Figure 2.76. By Equation 2.10 for

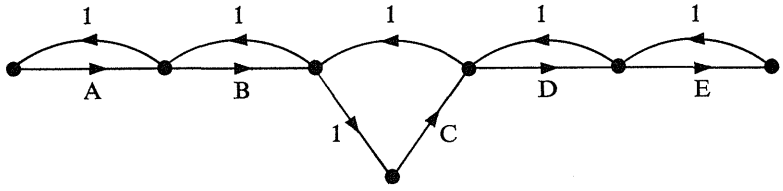


FIGURE 2.76

$\Delta^\dagger$, we may find the determinant $\Delta^\dagger$ of the new graph by adding on to Δ the additional terms:

$$-\sum_{\nu=1}^{k} L_\nu' \Delta_\nu$$

associated with the *new loops* that are formed. In this specific example, only *one* new loop, L_1' is created and its transmittance is C. The associated cofactor Δ_1 is the determinant of the subgraph which remains after erasing all parts of the graph that touch this loop, as shown in Figure 2.77. Since the subgraph consists of two parts, its determinant is given by the product of the determinants of each part.

$$\Delta_1 = (1 - A)(1 - E).$$

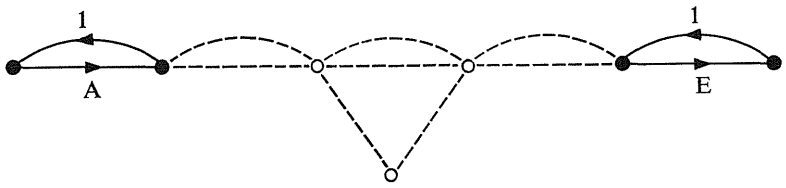


FIGURE 2.77

Hence, by Equation 2.10,

$$\Delta^{\dagger} = \Delta - L_1'\Delta_1$$

$$= (1 - A - B)(1 - D - E) - C(1 - A)(1 - E).$$

You will find it instructive to show that this expression for $\Delta^{\dagger}$ is actually equal to the determinant for the graph in Question 2.25 given on page 91, even though this expression contains only 7 symbols while the other contains 20. This comparison suggests that temporary removal of a branch or node from a graph may greatly simplify the calculation of the graph determinant; one first evaluates the determinant in the absence of some branch or node and then adds on the additional terms that arise when the branch or node is replaced.

At last, we are in a position to give Mason's expression for the graph transmittance between a specified source and sink. You will recall that the graph transmittance G_{jk} is equal to the signal appearing at the node k per unit of signal injected at node j. This may best be visualized by adding to the graph a source node j' and sink node k using unit transmittances, as illustrated in Figure 2.78. G_{jk} is then the transmittance of the single

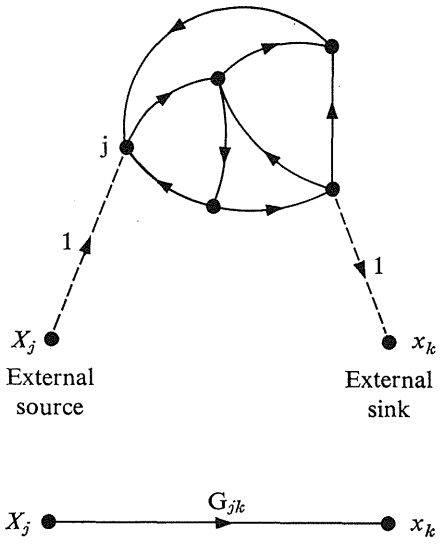
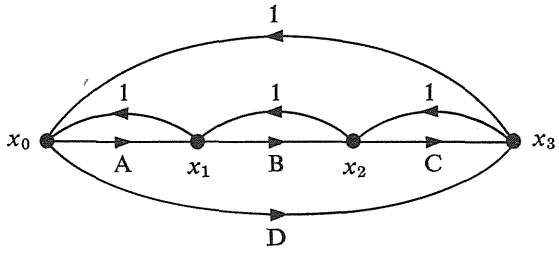


FIGURE 2.78

QUESTION 2.41 By temporarily removing the node x_0 from the following graph, evaluate its determinant, using the method described in the previous paragraph. (Answer)



branch obtained by absorbing all other nodes. We now derive an alternative expression for any graph transmittance G_{jk} in terms of nontouching loops and paths of the graph.

Let us suppose that the original graph contains n nodes. Then, we may regard the source and sink nodes used in defining the graph transmittances as the two halves of the split $(n + 1)$st node as sketched in Figure 2.79. Next, let T be the loop transmission

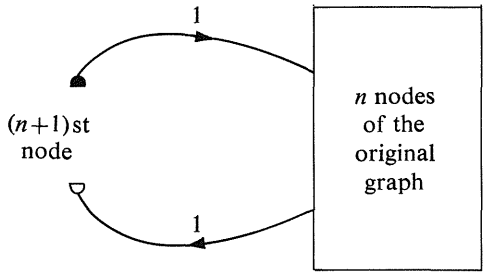


FIGURE 2.79

of node $n + 1$. Since $T = T'_{n+1}$, if we denote by Δ the determinant of the n-node graph, the determinant $\Delta^\dagger$ of the new graph formed by incorporating node $n + 1$, will be by Equation 2.7:

$$\Delta^\dagger = \Delta(1 - T).$$

Hence,

$$1 - T = \frac{\Delta^\dagger}{\Delta}.$$

Furthermore, by Equation 2.10 we have

$$\frac{\Delta^\dagger}{\Delta} = 1 - \frac{\sum_{v=1}^{k} L'_v \Delta_v}{\Delta}.$$

Hence,

$$T = \frac{\sum_{v=1}^{k} L_v \Delta_v}{\Delta}.$$

But if node $n + 1$ is permanently split, T is just the source-to-sink transmittance of the graph and the L'_ν is the transmittance of the νth source-to-sink path. We shall therefore introduce a slightly different notation for these quantities, in keeping with the new interpretation:

G = source-to-sink graph transmittance (the sink signal per unit of source signal, identical to T),

P_ν = transmittance of the νth source-to-sink open path (identical to L'_ν),

Δ = determinant of original graph,

Δ_ν = cofactor of the νth path (the determinant of that part of the graph not touching the kth path).

Thus, we have the important result:

Mason's General Graph-Transmittance Expression

$$G = \frac{\sum_\nu P_\nu \Delta_\nu}{\Delta}$$

$$= \left[\frac{(P_1 + P_2 + \cdots + P_k)(1 - L_1)(1 - L_2)\ldots(1 - L_m)}{(1 - L_1)(1 - L_2)\ldots(1 - L_m)}\right]^*, \qquad (2.6)$$

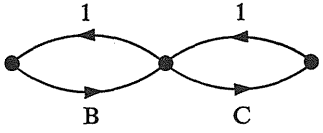
where * indicates that any terms containing transmittance products of confluent branches are to be dropped.

The cofactor Δ_ν is formed by deleting from Δ any terms which contain branches touching the path P_ν. This cofactor may also be found by first erasing all parts of the original graph that touch the path P_ν and then evaluating the determinant of the subgraph that remains. Either method will yield the same result.

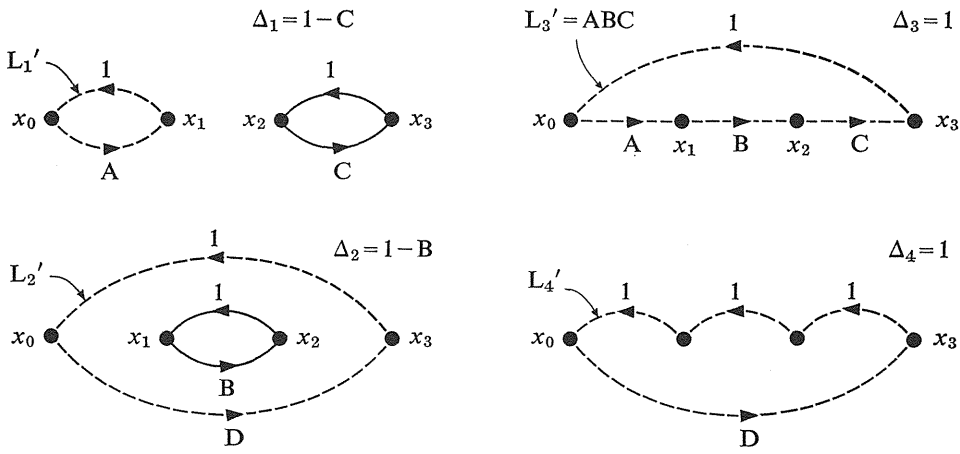
ANSWER TO QUESTION 2.41

$$\Delta = (1 - B - C) - A(1 - C) - D(1 - B) - D - ABC.$$

When node x_0 is removed (and all branches touching it), the determinant of the remaining graph is



When x_0 is replaced, four new loops are formed: $L'_1 = A$, $L'_2 = D$, $L'_3 = ABC$, $L'_4 = D$. The corresponding values of Δ_1 through Δ_4 are the determinants of the subgraphs which do not touch these paths.



Hence,

$$\Delta^\dagger = \Delta - L_1'\Delta_1 - L_2'\Delta_2 - L_3'\Delta_3 - L_4'\Delta_4$$
$$= (1 - B - C) - A(1 - C) - D(1 - B) - ABC - D.$$

Appendix B: References in Chronological Order

1953 1. Mason, S. J., "Feedback Theory—Some Properties of Signal Flow Graphs." *Proceedings of the IRE, 41*, No. 9 (Sept. 1953).

1955 2. Truxal, J. G. *Automatic Feedback Control System Synthesis.* New York: McGraw-Hill.

1956 3. Mason, S. J., "Feedback Theory—Further Properties of Signal Flow Graphs." *Proceedings of the IRE, 44*, No. 7 (July 1956).

1957 4. Putterman, H., "Signal Flow Graph Technique." *Sperry Engineering Review, 10*, No. 6 (Nov.–Dec. 1957).

1959 5. Cheng, D. K. *Analysis of Linear Systems.* Reading, Mass: Addison-Wesley.

6. Seshu, S., and Balabanian, N. *Linear Network Analysis.* New York: Wiley.

7. Warfield, J. N. *Introduction to Electronic Analog Computers.* Englewood Cliffs, N.J.: Prentice-Hall.

8. Coates, C. L., "Flow Graph Solutions of Linear Algebraic Equations." *IRE Transactions on Circuit Theory, CT-6*, No. 2 (June 1959).

9. Hoskins, R. F., "Signal Flow Graphs." *Electronic and Radio Engineer* (British), *36*, No. 8 (Aug. 1959).

10. Kauffmann, P., and Klein, J., "Flow Graph Analysis of Transistor Feedback Networks." *Semiconductor Products*, Oct. 1959.

11. Dunn, S. B., and Ross, G. F., "Signal Flow and Scattering Technique in Microwave Network Analysis." *Sperry Engineering Review, 12*, No. 4 (Dec. 1959).

12. Happ, W. W., and Nisbet, T. R., "Visual Engineering Mathematics." *Electronic Design, 7*, No. 25 (Dec. 9, 1959); *7*, No. 26 (Dec. 23, 1959);

1960 *8*, No. 1 (Jan 6, 1960); *8*, No. 2 (Jan. 20, 1960).

13. Mason S. J., and Zimmerman H. J. *Electronic Circuits, Signals, and Systems.* New York: Wiley. (This is an outstanding text and reference which was used extensively in preparing this book.)

1961 14. Seshu, S., and Reed, M. B. *Linear Graphs and Electrical Networks.* Reading, Mass: Addison-Wesley.

15. Balabanian, N., *Circuit Theory.* Boston: Allyn and Bacon, Chapter 11.

16. Lynch, W. A., and Truxal, J. G. *Introductory System Analysis.* New York: McGraw-Hill, Chapter 7.

1962 17. Chow, Y., and Cassignol, E. *Linear Signal-Flow Graphs and Applications.* New York: Wiley.

18. Robichaud, L. P. A., Boisvert, M., and Robert, Jean. *Signal Flow Graphs and Applications.* Englewood Cliffs, N.J.: Prentice-Hall.

1963 19. Ore, O. *Graphs and Their Uses.* New Mathematical Library, Vol. 10, Random House, New York. (This is an inexpensive paperback that offers a beautifully lucid account of linear-graph theory with its manifold applications to physical problems.)

1964 20. Lorens, C. S., "Flowgraphs." McGraw-Hill Monographs in Modern Engineering Science, 1964. (This inexpensive paperback of 178 pages is a lucid, stimulating explanation of flow-graph techniques to a wide variety of problems. It provides very useful supplementary reading to this textbook and contains many references on flow-graph applications.)

Summary—Definitions of Terms*

Branch A line segment joining two *nodes*, or joining one *node* to itself.

Branch Input Signal The signal x_j, at the input end of branch *jk*.

Branch Operator The operation performed on the *branch input signal* to produce the *branch output signal.*

Branch Output Signal (of branch *jk*) The component of signal x_k contributed by node *j* via branch *jk*.

Branch Transmittance The ratio of *branch output signal* to *branch input signal.*

Cascade Node (*Branch*) A *node* (*branch*) not contained in a *loop.*

Cofactor (*or Path Cofactor*) See *Path* (*Loop*) *Factor.*

Dependent Node A *node* having one or more incoming *branches.*

Directed Branch A *branch* having an assigned direction.

NOTE: In identifying the branch direction, the branch *jk* may be thought of as *outgoing* from node *j* and *incoming* at node *k*. Alternatively, branch *jk* may be thought of as *originating* or having its *input* at node *j*, and *terminating* or having its *output* at node *k*. The assigned direction is conveniently indicated by an arrow pointing from node *j* toward node *k*.

Feedback Node (*Branch*) A *node* (*branch*) contained in a *loop.*

* *IRE Proceedings*, *48* (Sept. 1960), p. 1611.

Graph Determinant One plus the sum of the *loop-set transmittances* (*operators*) of all *nontouching loop sets* contained in the graph. (Note that by the definition of the *loop-set* transmittances, some of these terms carry a negative algebraic sign.)

[NOTE 1: The graph determinant is conveniently expressed in the form:

$$\Delta = \left[1 - \sum L_i + \sum L_i L_j - \sum L_i L_j L_k + \cdots\right]^*, \tag{2.5}$$

where L_i represents all the different loops of the graph, $L_i L_j$ represents all the different pairs of nontouching loops, $L_i L_j L_k$, all the different nontouching triplets, etc.]

[NOTE 2: The graph determinant may be written alternatively as

$$\Delta = \left[(1 - L_1)(1 - L_2), \ldots, (1 - L_n)\right]^*, \tag{2.4}$$

where $L_1, L_2, \ldots, L_n$ are the loop transmittances (operators) of the n different loops in the graph, and where the asterisk indicates that, after carrying out the "multiplications" within the brackets, a term will be dropped if it contains the transmittance (operator) product of two touching loops.]

[NOTE 3: The *graph determinant* reduces to the *return difference* for a graph having only one loop.]

[NOTE 4: The graph determinant is equal to the determinant of the coefficient equations.]

Graph Transmittance (*Operator*) The equivalent operator that yields the signal at some specified *dependent node*, when applied to the signal at some specified *source node*.

[NOTE: The graph transmittance (operator) is the weighted sum of the path transmittances (operators) of the different *open paths* from the designated *source node* to the designated *dependent node*, where the weight for each path is the *path factor* divided by the *graph determinant*.]

Loop (*Feedback Loop*) A simple closed *path* in which no node is traversed more than once.

Loop Graph A *signal flow-graph* each of whose branches is contained in at least one *loop*.

[NOTE: Any *loop* graph embedded in a general graph can be found by removing the *cascade branches*.]

Loop-Set Transmittance The product of the *negatives* of the transmittances of the *loops* in a set.

Loop Transmittance (*Operator*) The product of the *branch transmittances* (*operators*) of a *loop*.

Loop Transmittance of a Branch The loop transmittance of an interior node inserted in that branch.

[NOTE: A branch may always be replaced by an equivalent sequence of *branches*, thereby creating interior *nodes*.]

Loop Transmittance of a Node The graph transmittance from the *source node* to the *sink node* created by splitting the designated *node*.

Node One of the set of discrete points in a flow-graph.

Node Absorption A flow-graph transformation whereby one or more *dependent nodes* disappear and the resulting graph is equivalent with respect to the remaining *node signals.*

[NOTE: For example, a circuit analog of *node absorption* is the star-delta transformation.]

Node Signal A variable, x_k, associated with node k.

Nontouching Loop Set A set of loops no two of which have a common *node.*

Open Path A *path* along which no *node* appears more than once.

Path Any continuous succession of *branches,* traversed in the indicated branch directions.

Path (Loop) Factor The *graph determinant* of that part of the graph not touching the specified *path (loop).*

[NOTE 1: A *path (loop) factor* is obtainable from the *graph determinant* by striking out all terms containing transmittance products of loops which touch that *path (loop).*]

[NOTE 2: For loop L_k, the *loop factor* is $-\partial\Delta/\partial L_k$.]

Path Transmittance (Operator) The product of the *branch transmittances (operators)* in that *path.*

Return Difference Unity minus the *loop transmittance.*

Signal Flow-Graph A network of *directed branches* in which each dependent *node signal* is the algebraic sum of the incoming *branch signals* at that *node.*

[NOTE: Thus, $x_1t_{1k} + x_2t_{2k} + \cdots + x_nt_{nk} = x_k$, at each dependent node k, where t_{jk} is the *branch transmittance* of branch jk.]

Sink Node A node having only incoming *branches.*

Source Node A *node* having only outgoing *branches.*

Split Node A node that has been separated into a *source node* and a *sink node.*

[NOTE 1: Splitting a *node* interrupts all signal transmission through that *node.*]

[NOTE 2: In splitting a node, all incoming branches are associated with the resulting *sink node,* and all outgoing branches with the resulting *source node.*]

In this chapter we have concentrated on learning flow-graph rules and transformations without attempting to show how signal flow-graphs may be used to study physical systems. The methods that you have learned are extremely powerful and useful. In the following chapters of this book you will have an opportunity to use these same methods in studying a variety of different physical systems. In particular, we shall use these methods in the next chapter to study signals in electrical circuits formed of resistors, vacuum tubes, and transistors, and we shall develop the theory of the operational amplifier which is the essential element in analog computers.

4 Operator Graphs

You will recall from the discussion in Chapter 1 that an operator may be classified as *static* or *dynamic*, *linear* or *nonlinear*, *stationary* or *time-varying* according to how it transforms any signal x into another signal y. You might think that a very large number of different kinds of operators would be needed to describe the signal relationships in physical systems. Fortunately, this is not so. In fact, we can do quite nicely with just three basic operators: *scalors*, *delayors*, and *limitors*. As we shall see, we require only *two* kinds of *linear* operators and *one* kind of *nonlinear* operator as fundamental "building blocks" to represent practically all physical operations of interest to us.

Each of these three basic operators may be realized physically in many different ways. Sometimes a single device may be used to perform several operations simultaneously; at other times a fairly elaborate circuit of many physical elements may be needed to realize a single operation. For instance, an operational amplifier involving several transistors, resistors, capacitors, and power sources may be used to construct an accurate scalor. Likewise, a very complicated circuit involving a multitude of electronic components, including perhaps even a magnetic tape recorder, may be required to construct a delayor. In this book, we shall focus attention on the operator rather than on the physical device, because devices for performing these different operations not only are of unlimited variety, but, furthermore are subject to rapid change and obsolescence. We use only three basic operators which will never become obsolete and will be useful always. By following this approach, we can ignore, temporarily, the peculiarities of particular physical devices. To illustrate engineering design by studying specialized physical systems opens a Pandora's box that is chock-full of complications, many of which are quite irrelevant to the operational aspect of the system. Even though these complications are of practical importance in the final execution of the design of a *specific* physical system, the *operational aspect* of the system is also meaningful by itself.

The operational aspect can be studied separately from the physical aspects that pertain only to a particular realization. This is done by dealing with "pure" systems constructed from the three basic operators just listed. In this axiomatic approach to systems, the elements are exactly defined, and we have explicit rules for combining the elements. These pure systems are thus mathematical structures which are exactly what we

define them to be and nothing more. They illustrate some of the important steps in the engineering design process. In this, we are doing engineering with generalized mathematical entities, rather than with specialized physical entities. And by using the digital or analog computer, we may study these pure systems and observe their performance in the most intimate detail.

Classification of Operators

Before proceeding, we must have a clear understanding of what is meant by the adjectives *static*, *dynamic*, *linear*, *nonlinear*, *stationary*, and *time-varying*. The following discussion should help in clarifying the physical significance of these properties, first discussed in Chapter 1.

A small electric-clock motor is connected, as shown in Figure 4.1, through a gear box and a mechanical linkage to the movable arm of an adjustable resistance attenuator of the

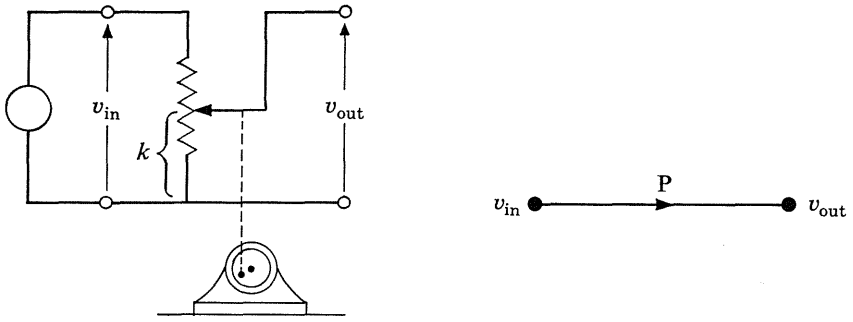


FIGURE 4.1 *A motor-driven adjustable attenuator.*

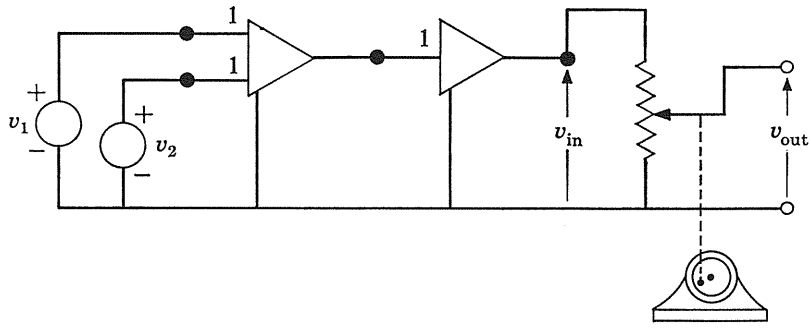
sort discussed in Chapter 3. The motor runs at a constant speed and has been set so that the position of the moving contact k varies with the time t in accord with the accompanying equation,

$$k(t) = 0.5\left[1 - \cos\left(\frac{\pi t}{30}\right)\right]$$

where k is the fraction of the total resistance between the moving contact and ground, and t is the time in seconds after midnight. The aspect of the system in which we are interested is the operator P, which relates the input voltage v_{in} to the output voltage v_{out}.

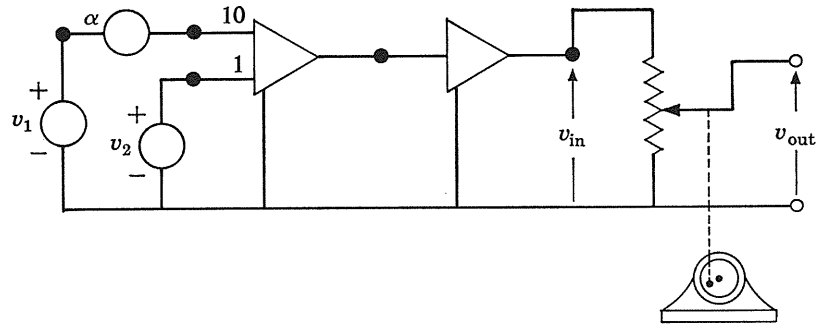
QUESTION 4.1 If the value of the input voltage of Figure 4.1 at any instant t is given by the function $v_{in}(t)$, what is the value $v_{out}(t)$ of the output voltage at this same instant? (Answer)

QUESTION 4.2 Suppose that the input voltage v_{in} of Figure 4.1 is composed of the sum of two voltages v_1 and v_2.



If the values of v_1 and v_2 at any instant t are $v_1(t)$ and $v_2(t)$, respectively, what is the value of v_{in} and v_{out} at the instant t ? (Answer)

QUESTION 4.3 Suppose that instead of v_1, a different signal 10α times larger than v_1 had been used.



What is the equation that expresses $v_{out}(t)$ in terms of $v_1(t)$ and $v_2(t)$? (Answer)

QUESTION 4.4 Is the operator P that relates v_{in} to v_{out} *static* or *dynamic*? (Answer)

QUESTION 4.5 Remember that an operator is *linear* if:

$$(1)\quad [f_1 + f_2]\mathrm{L} = f_1\mathrm{L} + f_2\mathrm{L},$$

$$(2)\quad (af_1)\mathrm{L} = a(f_1\mathrm{L}).$$

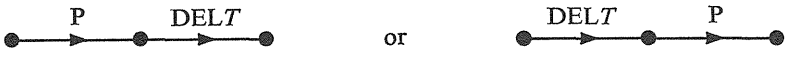
Is the operator P linear? (Answer)

QUESTION 4.6 A delayor is an operator DELT which operates on any signal f, whose value at the instant t is $f(t)$, and produces a signal whose value at this *same* instant is $f(t - T)$:

where $f\mathrm{DEL}T(t) = f(t - T)$.

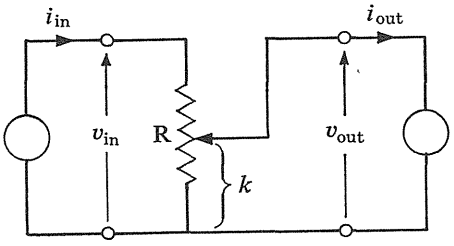
Note that the value $f(t - T)$ of the output signal at any instant t is the value that the input signal f assumed at the instant T time units prior to t.

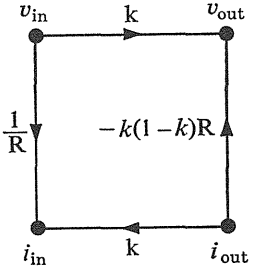
You will recall from Chapter 1 that an operator is *stationary* if it commutes with any delayor. If the operator P is stationary, then the same overall operation is achieved with either



Is the operator P relating v_{in} to v_{out} stationary? (Answer)

ANSWER TO QUESTION 4.1 According to the results of the previous chapter, the input and output signals in a resistance attenuator may be described by the flow-graph shown at the right. For any resistance the values of the voltage and current





signals are directly proportional at each instant and do *not* depend on the signal values at other instants. Hence, if the position of the adjustable contact at any instant t is denoted by the function $k(t)$, the value of the output voltage v_{out} at that instant may be expressed in terms of the values $v_{in}(t)$ and $i_{out}(t)$ as

$$k(t)\cdot v_{in}(t) - k(t)[1 - k(t)]Ri_{out}(t) = v_{out}(t). \tag{4.1}$$

If the output is *open-circuited* so that $i_{out}(t) = 0$, we have

$$k(t)v_{in}(t) = v_{out}(t). \tag{4.2}$$

Then, at each instant the value of the output voltage is proportional to the value of the input voltage at that *same* instant. However, the coefficient of proportionality is different at different instants:

$$0.5\left[1 - \cos\left(\frac{\pi t}{30}\right)\right]v_{in}(t) = v_{out}(t).$$

ANSWER TO QUESTION 4.2 If $v_1 + v_2 = v_{in}$, then at any instant

$$v_1(t) + v_2(t) = v_{in}(t)$$

and

$$k(t)\cdot v_{in}(t) = v_{out}(t)$$

or

$$k(t)v_1(t) + k(t)v_2(t) = v_{out}(t).$$

ANSWER TO QUESTION 4.3 If $10\alpha v_1 + v_2 = v_{in}$, then, at any instant,

$$10\alpha \cdot v_1(t) + v_2(t) = v_{in}(t).$$

Since

$$k(t)v_{in}(t) = v_{out}(t),$$

$$10\alpha \cdot k(t)v_1(t) + k(t)v_2(t) = v_{out}(t).$$

That is, the particular term in $v_{out}(t)$, associated with v_1, is simply multiplied by 10α as a consequence of increasing v_1 by the scalar factor 10α.

ANSWER TO QUESTION 4.4 Since the value of the output signal v_{out} at any instant depends only on the value of the input signal v_{in} at the *same* instant, the operator is *static*.

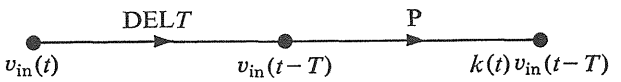
ANSWER TO QUESTION 4.5 The operator P is *linear* since, as shown in the answer to Question 4.2,

$$(v_1 + v_2)\mathrm{P} = v_1\mathrm{P} + v_2\mathrm{P}$$

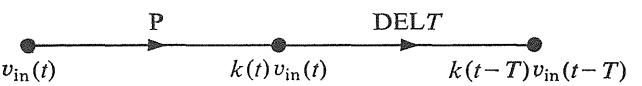
and in the answer to Question 4.3,

$$[\alpha v_1]\mathrm{P} = \alpha[v_1\mathrm{P}].$$

ANSWER TO QUESTION 4.6 If the operator P is stationary, the result v_{in}DELT of operating on the delayed input signal should be the same as the result of delaying v_{in}P. In the first case, the values of the signals at any instant t are



In the second case, the signal values are

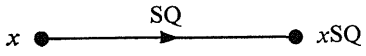


But, in general, for any time delay T

$$k(t)v_{in}(t - T) \neq k(t - T)v_{in}(t - T)$$

since $k(t) \neq k(t - T)$. Hence, the operator P is *nonstationary* (i.e., time-varying). Only when the value of $k(t)$ is a constant that does not change with time will these expressions be equal.

The operator discussed above is linear, static, but nonstationary. All the various combinations of these properties are found in physical systems. For instance, a squaring operator, SQ, is available in many analog computers. It operates on any signal x to yield another signal xSQ, the value of which at any instant is the square of the value of the input signal at that instant:



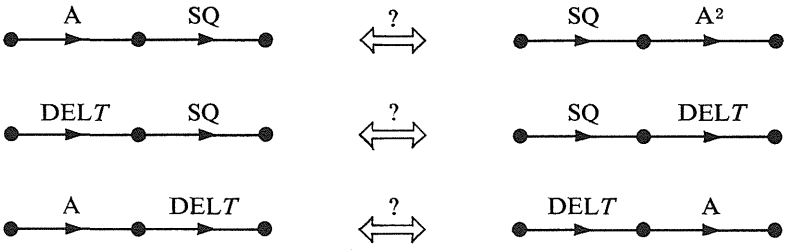
where $x\text{SQ}(t) = [x(t)]^2$.

FIGURE 4.2 *A squaring operator.*

This operator is *static* because the value of the output at any instant depends only on the value of the input at that *same* instant.

QUESTION 4.7 Is the operator SQ linear? Is it stationary? (Verify your answers by determining whether SQ satisfies the definitions of linearity and stationarity.) (Answer)

QUESTION 4.8 Let A be a scalor, SQ a squaring operator, and DELT a delayor, of delay time T. Which of the following pairs of cascaded operators yield equivalent transformations of the input signal? (Answer)



Systems of Operators

In Chapter 2 we learned how to interpret signal flow-graphs and how to relate them to the conventional algebraic symbols. The rules were:

1. The signals (i.e., observables) are denoted by nodes.
2. The additive effect of one signal on another is denoted by a *directed branch* from the *causal* node to the *dependent* node.
3. The signal at any *dependent* node is composed of the total signal flow *into* that node from all *incoming* branches. (The signal at each *source* node is assumed to be specified independently.)
4. The signal at a node is operated upon by each outgoing branch. (That is, outgoing branches in no way affect the signal at a node.)

5. The signal at the input end of a branch is transformed or modified in some specified way as it passes through the branch. The branch therefore represents a prescribed *operation* upon the input signal to obtain the transformed signal at its output. Thus, in general, the branch denotes an *operator*.

These various conventions are illustrated by the simple graphs shown in Figure 4.3. You should examine these carefully and thoroughly understand their significance. In particular, observe that no restriction whatsoever is placed on the operators P, Q, and R. They may be *nonlinear* and *time-varying* without violating the flow-graph rules just given.

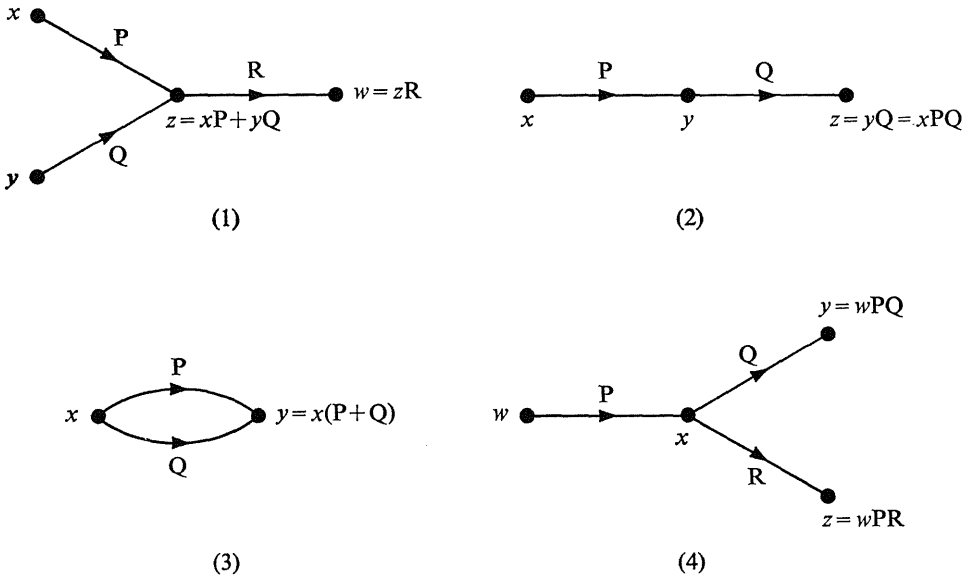
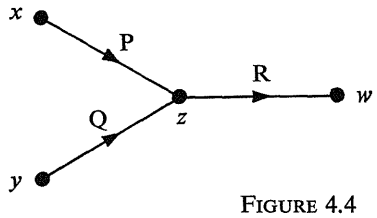


FIGURE 4.3 *Simple operator graphs.*

In the graph of Figure 4.3(1), the signal z is shown to depend on the signals x and y:



$xP + yQ = z.$

FIGURE 4.4

Here we follow the convention of writing the operator symbol on the *right* of the signal on which it acts. Hence, the signal z is *composed* of two component signals, xP and yQ, which are the results respectively of the operation P on x and the operation Q on y. By the same convention, $w = z$R denotes the signal w created as a result of the operation R on the signal z.

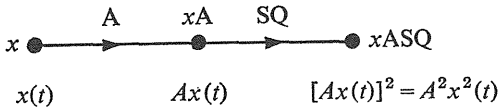
Thus far, we have not attached any numbers or numerical values to our signals. We shall assume that the signals are completely described if we specify their values for each

instant of time t. We shall use the notation $z(t)$, $x\mathrm{P}(t)$, $y\mathrm{Q}(t)$, $z\mathrm{R}(t)$, and $w(t)$, to denote the *value* at the same instant t of each of the signals considered thus far. Note particularly that the combination of two or more letters such as $x\mathrm{P}$ is interpreted as a *single name* for a signal and does *not* imply arithmetic multiplication. (Computer languages, such as FORTRAN, also name the functions of variables by grouping letters together in this way.) An additive expression, such as $x\mathrm{P} + y\mathrm{Q} = z$, implies that the values of z, $x\mathrm{P}$, and $y\mathrm{Q}$ must satisfy the numerical equation

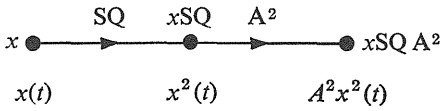
$$z(t) = x\mathrm{P}(t) + y\mathrm{Q}(t)$$

for *all* values of t. Likewise, $w(t) = z\mathrm{R}(t)$.

ANSWERS TO QUESTIONS 4.7 AND 4.8 A linear operator L must necessarily commute with a scalor: AL = LA. Consider now the effect of the operator sequence A SQ on a signal x. The values of the resulting signals at any instant are

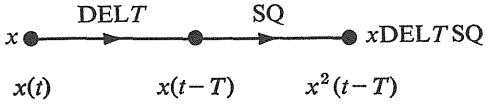


This evidently yields the same end value as the operator sequence, SQ A^2,

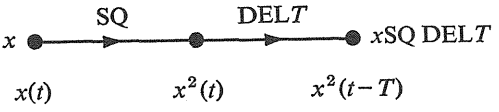


Evidently ASQ $\neq$ SQA and, hence, SQ is *not* a linear operator.

If an operator commutes with any delayor, it is stationary. Now for the operator sequence, DELT SQ, we have the signal values

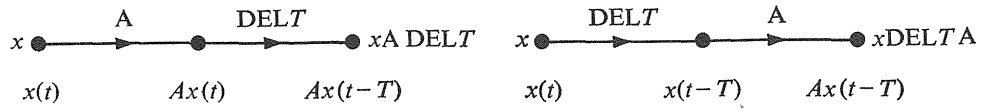


whereas the sequence SQ DELT yields



Evidently, DELT SQ = SQ DELT and, hence, SQ is a *stationary* operator.

Similarly, the delayor and scalor commute because



Hence, the scalor is a *stationary* operator.

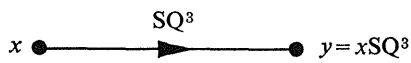
The operators P, Q, and R may be of a very general kind and still satisfy perfectly the basic flow-graph definitions. For instance, in graph 2 of Figure 4.3, $z = x\mathrm{PQ}$; in graph 3, $y = x\mathrm{P} + x\mathrm{Q}$; and in graph 4, $y = w\mathrm{PQ}$ and $z = w\mathrm{PR}$. But unless we specialize them to the much narrower class of linear operators, we shall be unable to use the algebraic reduction rules for manipulating, reducing, and solving the graph.

Repeated operations • When we begin to combine several operators into a single equivalent operator, we need further refinement of our algebraic notation. We have consistently tried to distinguish among a *signal*, an *operator*, and a *function*. An operator acts on a signal to give another signal. The signals may be *described* by functions of time to which are assigned the usual properties of continuity, differentiability, etc. The concept of *operator* is associated with the physical entity itself, and is more fundamental than the numerical functions which *we* introduce for purposes of analysis. Because of similarities in notation it is rather easy to get functions and operators confused, particularly with respect to rules for exponentiation. In this section, we shall clarify these rules and also indicate a procedure for realizing the inverse of a given operator.

Repeated application of an operator is shown by an exponent attached to the operator itself. Thus, for any operator P, the result of applying the operator n times to a signal x may be written as $x\mathrm{P}^n$. In flow-graph notation, for $n = 3$,

FIGURE 4.5

For instance, if P were the squarer SQ



where $y(t) = x\mathrm{SQ}^3(t)$
$= [x(t)]^8$.

FIGURE 4.6

Note that this is altogether *different* from the use of an exponent to denote raising the *value* of some function to a power:

$$(x\mathrm{SQ}(t))^3 = (x^2(t))^3 = x^6(t)$$

In the case of the repeated squaring operation,

$$y = x\ \mathrm{SQ}\ \mathrm{SQ}\ \mathrm{SQ} = x^2\mathrm{SQ}\ \mathrm{SQ}$$
$$= x^4\mathrm{SQ} = x^8,$$

where the value of the signal x^8 is *defined* to be $[x(t)]^8$ at the instant t, by definition, $x^8(t) \equiv [x(t)]^8$, so that $y(t) = x^8(t)$. Similarly, for an exponentiation operator EXP, where $x\mathrm{EXP}(t) = e^{x(t)}$,

where $y(t) = (e)^{e^{x(t)}}$

FIGURE 4.7

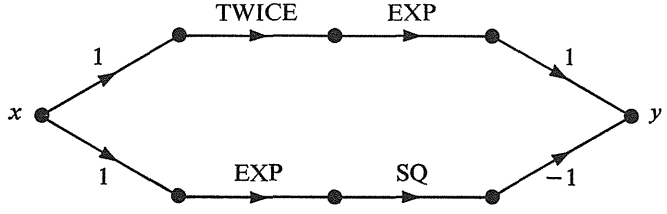
rather than $y(t) = [\exp x(t)]^2$.

QUESTION 4.9 Using only the operators thus far introduced, how would you specify the operation on a signal x which will yield a signal y whose value at any instant is $y(t) = [\exp x(t)]^2$? (Answer)

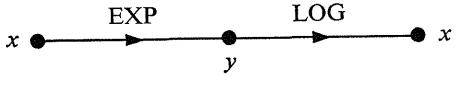
QUESTION 4.10 Consider the operator, TWICE, which in operating on any signal x will yield a signal xTWICE whose value at any instant t is

$$x\text{TWICE}(t) = 2x(t).$$

In terms of the function $x(t)$, what is the value at any instant of the output signal $y(t)$ shown in the accompanying graph? (Answer)



Inverse of an operator • In keeping with the exponentiation convention just described, it is reasonable to denote the *inverse* of P by the symbol P^{-1} (provided such an inverse exists). For instance, if P is the operator EXP, the inverse operator P^{-1} would be LOG, since



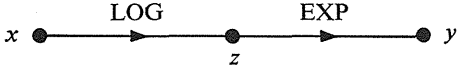
$x\text{EXP LOG} = x.$

FIGURE 4.8

Hence, we may denote LOG by EXP^{-1}, and the two operations taken together are equivalent to the identity operator. This relationship will be valid for a signal that takes on any value whatever, either positive or negative. However, it should be noted, that y *never* takes on a *negative* value:

$$y\text{LOG}(t) \underset{(y(t)>0)}{=} \log [y(t)] = x(t).$$

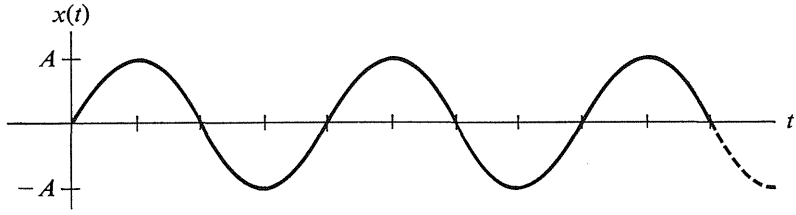
QUESTION 4.11 What constraint must be imposed on the values of the signal x for the following sequence of operations to be equivalent to the *identity operator*? (that is, for $y = x$) (Answer)



QUESTION 4.12 Another important nonlinear static operator is the absolute-value operator ABS, defined for any x as

where $x\text{ABS}(t) = |x(t)|$.

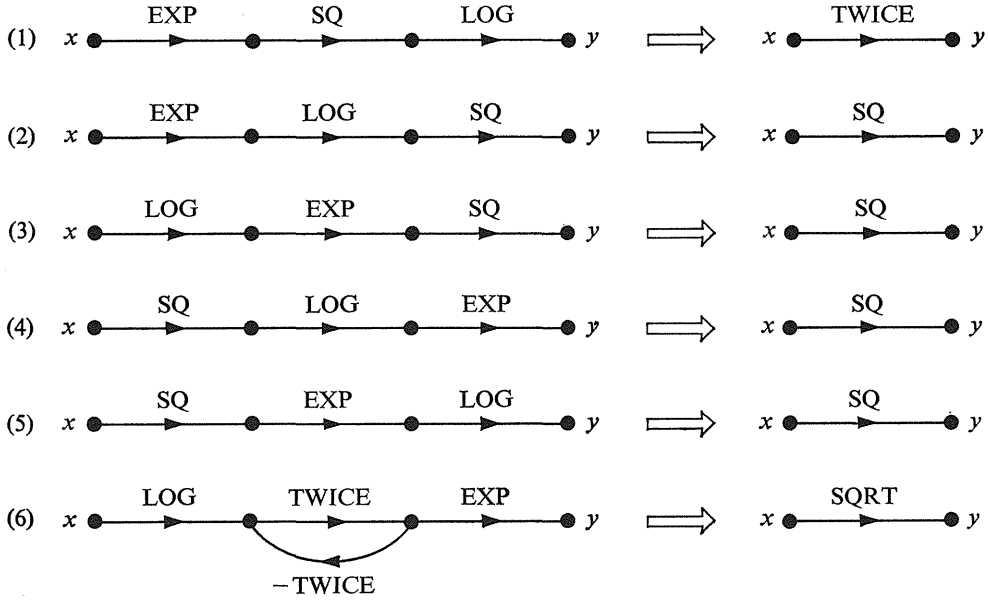
Suppose that x is a sinusoidally varying signal, described by the function $x(t)$ plotted below.



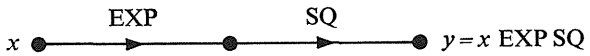
Plot the function $x\text{ABS}(t)$. (Answer)

QUESTION 4.13 What single operator previously considered is equivalent to the following composite operator? (Answer)

QUESTION 4.14 For which of the graphs shown below are the implications correct for any x? (Answer)

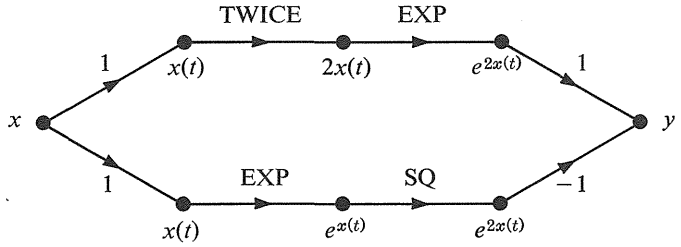


ANSWER TO QUESTION 4.9 To create a signal whose value is $[\exp x(t)]^2$ when the input signal has a value $x(t)$, we may first form a signal whose value is $\exp x(t)$ by applying the exponential operator EXP to x and then "squaring" this signal:



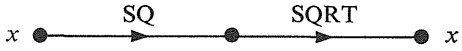
$$x\text{EXP SQ}(t) = [\exp x(t)]^2$$

ANSWER TO QUESTION 4.10 Consider the value of the signal at each node at any instant t,



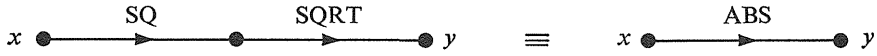
Evidently, xTWICE EXP = xEXP SQ so the value $y(t)$ of the output signal is zero at every instant for any $x(t)$.

ANSWER TO QUESTION 4.11 Observe that here we must restrict the *domain* of signals on which LOG operates to that class of signals which *do not have negative* values. For this class of *positive-valued* signals, EXP = LOG^{-1}. A similar difficulty arises with the following graph, involving a square-root operator SQRT. Here, also, provided x is the class of *positive-valued* signals,



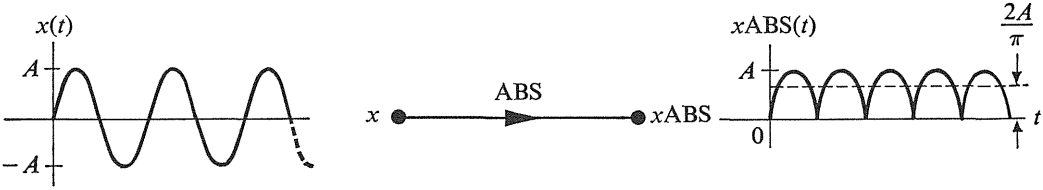
SQRT = SQ^{-1},

and the combination is equivalent to the identity operator. More generally, the combination is equivalent to the absolute-value operator, ABS, where for any input signal x, xABS$(t) = |x(t)|$:

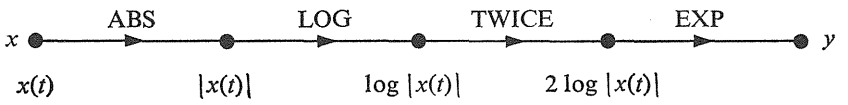


where $y(t) = x\text{ABS}(t) = |x(t)|$.

ANSWER TO QUESTION 4.12 The operator ABS describes a "full-wave rectifier" which is sometimes used in electronic circuits to convert an alternating voltage (having zero average value) to a unipolar voltage whose average value is $2/\pi$ times the peak value A of the sinusoidal signal,



ANSWER TO QUESTION 4.13 Consider the value of each signal in this graph at any instant t

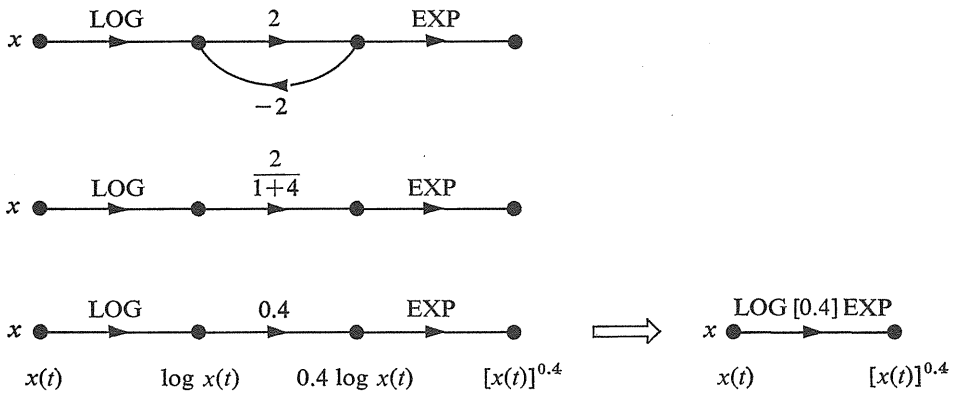


$$\therefore\ y(t) = e^{2\log|x(t)|} = |x(t)|^2 = x^2(t). \tag{4.3}$$

Hence, this operator is equivalent to the squaring operator, SQ. This equivalence holds for all real-valued functions $x(t)$.

ANSWER TO QUESTION 4.14 The value of the output signal from the EXP is non-negative for any input signal, as is also the output from the SQ operator. The first, second, fourth, and fifth operators may be applied to any real-valued signal, whereas the third and sixth operators must be applied only to *positive-valued* input signals. Within this restriction, the single operator shown in the right-hand column produces the same output signal y as the system of operators in the left-hand column, for the first five systems given.

In the last graph, we recognize that TWICE is a scalor of transmittance 2, so the scalor loop can be reduced:



The preceding problems show that there are serious questions to be examined when one attempts to define the inverse P^{-1} of some physical operator P. In some cases the inverse is nonphysical, as when P is the delayor DELT. Here, the inverse is the *nonphysical anticipator* DEL$-T$. In other instances, the inverse exists, provided the domain of signals is suitably restricted. In yet other instances, the inverse may not exist physically in a strict sense but it may be approximated as closely as we desire (as in the case of a differentiator). In the following section we discuss a useful method of realizing an approximation to P^{-1} provided we are given the operator P. This technique is useful for realizing approximations on the analog computer.

Approximate realization of inverse of an operator • Our objective is to construct the inverse operator P^{-1} when we are given the operator P. Since $y = x\mathrm{P}$, we need to invert the path in some way that does not alter the branch P, as shown in Figure 4.9. The

FIGURE 4.9

desired result can be achieved by first placing in parallel with the operator branch P another path consisting of a scalor B in cascade with a branch of unity transmittance. Obviously, as the value of the transmittance $B \to 0$, the contribution of this added branch to the output signal will vanish, and, in the limit when $B = 0$, we have simply the results shown in Figure 4.10.

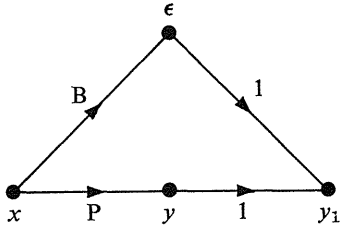
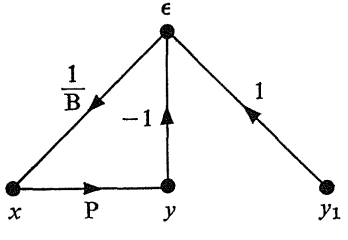


FIGURE 4.10

$$y_1 = x\mathrm{P} + x\mathrm{B}, \qquad \lim_{B \to 0} y_1 \Rightarrow y = x\mathrm{P}.$$

But now there are *two* paths from x to y_1. By inverting the *upper* path, we obtain the following graph, corresponding to the equations:



$$\epsilon = y_1 - y$$
$$x = \epsilon \mathrm{B}^{-1}$$
$$y = x\mathrm{P}$$

FIGURE 4.11

In this graph, the signal y_1 may be regarded as the *source* and x as the *dependent* signal. The *error* signal ϵ is the *discrepancy* between the signals y_1 and y. We have already seen that when B is made very small, the *error signal* $\epsilon = y_1 - y$ also becomes very small. Thus, we may approximately realize the inverse operator P^{-1} by using the equivalence shown in Figure 4.12. To achieve stable operation in practice the algebraic sign of the transmittance B^{-1} must be such that the signal x produces a change in $x\mathrm{P}$ tending to

reduce the error signal $\epsilon = y_1 - x\mathrm{P}$. The important point is that B may be an invertible operator (linear or nonlinear) that will yield a stable system. Detailed consideration of

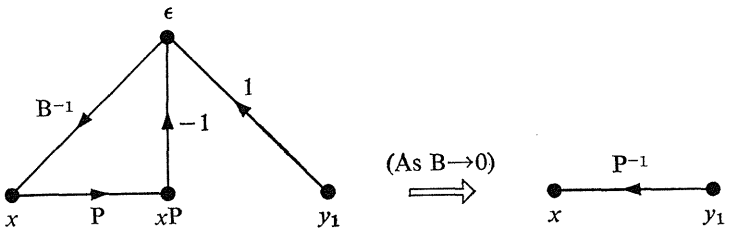
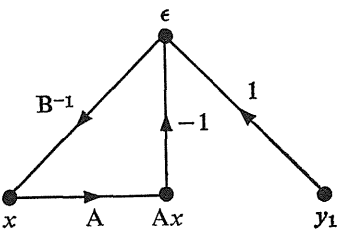


FIGURE 4.12 *Realization of* P^{-1} *from* P.

these important stability questions must be reserved until later, when we will have available the tools to answer them.

QUESTION 4.15 The method just described for obtaining the inverse of an operator P should certainly work when P is an ordinary scalor A. Suppose that A is a



scalor whose scalar transmittance is 10 and that y_1 is a constant-valued signal, $y_1(t) = 100$ for all t. What are the values of $\epsilon(t)$, $x(t)$, and $Ax(t)$ for each of the transmittance values of the scalor B? (Fill in the table.) (Answer)

B	$\epsilon(t)$	$x(t)$	$Ax(t)$
10.			
1.			
0.1			
0.01			
10^{-4}			

Multiplication and Modulation

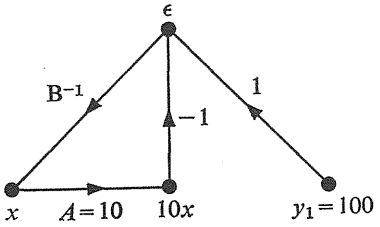
In these problems we have discussed several different nonlinear operators. The mathematical functions that describe these operators are perhaps so familiar to you that

you may not realize that building a physical device for performing the analogous mathematical operations is often quite difficult. For instance, to multiply two functions $f(t)$ and $g(t)$ *on paper*, you simply write $f(t)\cdot g(t)$. But to construct a physical device for performing this "multiplication" is not as easy as you might expect. Given two signals f and g, whose values at any instant are $f(t)$ and $g(t)$, we wish to construct a physical device having f and g as inputs which will yield an output signal r whose value at any instant is

$$f(t)\cdot g(t) = r(t).$$

This is quite a different process than manipulating symbolic marks on a piece of paper (although a digital computer does multiply by such symbolic manipulation). To emphasize this distinction, it may be of interest to mention a few of the schemes that have been invented for "multiplying" two signals.

ANSWER TO QUESTION 4.15 As this graph involves only scalors, the reduction methods of Chapter 2 may be applied to express all signals in terms of the source signal y_1:



$$\epsilon = y_1\left[\frac{1}{1 + AB^{-1}}\right],$$

$$\epsilon = y_1\left[\frac{B}{B + A}\right],$$

so that

$$\epsilon(t) = \frac{100\,B}{B + 10}.$$

Also,

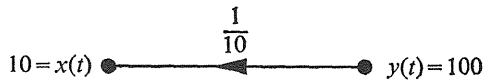
$$x = \epsilon B^{-1} = y_1\left[\frac{1}{B + A}\right],$$

so that

$$x(t) = \frac{100}{B + 10}.$$

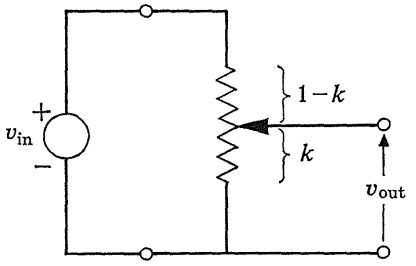
For various transmittance values of the scalor B, these expressions yield

B	$\epsilon(t)$	$x(t)$	$Ax(t)$
10	50	5	50
1	9.0909	9.0909	90.9090
0.1	0.9901	9.9010	99.0099
0.01	0.0999	9.9900	99.9001
0+	0+	10.−	100.−



Evidently, as $B \to 0$, the value of the signal x approaches 10, and hence the equivalent transmittance is $\frac{1}{10}$, which is clearly A^{-1}.

Servomultiplier • One of the most accurate schemes for "multiplying" two signals whose values are changing slowly as a function of time makes use of a variable resistance of the sort encountered earlier in Questions 4.1 and 4.2. The open-circuit voltage of the voltage divider was shown to be proportional to the applied voltage:



$$v_{out}(t) = k(t) \cdot v_{in}(t).$$

FIGURE 4.13

If we could make the applied voltage, v_{in}, proportional to the signal f, and the position, k, of the moving contact, proportional to the signal g, then it follows that v_{out} will be proportional to the signal r:

$$r(t) = cg(t)f(t),$$

when c is some scale factor. Several problems must be solved to use this scheme.

First, we must find some way to adjust the moving contact so that its position will be proportional to the value of the signal g at each instant. This clearly requires that the value of g must change sufficiently slowly that the mechanical motion can follow with negligible error. Evidently, discontinuous changes in the value of g could not be followed because physical mass and inertia will limit how rapidly the slider can be moved from one position to another.

Second, in the simple arrangement considered, the value of $g(t) = k(t)$ cannot be negative, although there is no restriction on the polarity of the input voltage, $f(t) = v_{in}(t)$. This arrangement is sometimes called a *two-quadrant multiplier* because if we describe the two input signals by a point with the value $g(t) = k(t)$ as the abscissa and the value $f(t) = v_{in}(t)$ as the ordinate, this point must lie in the first and fourth quadrants. (Actually, since $k(t)$ can never exceed 1, the plotted point must lie inside the strip illustrated by the shaded region in Figure 4.14.)

Let us first discuss a way to permit full *four-quadrant* operation. Consider the arrangement shown in Figure 4.15, which utilizes an operational amplifier as a scalor having a transmittance of -1. In this arrangement, both ends of the adjustable resistance

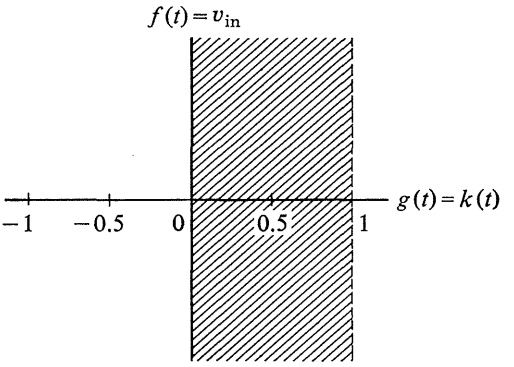


FIGURE 4.14

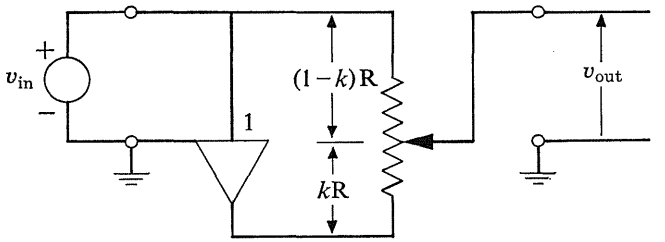


FIGURE 4.15

(potentiometer) have voltages applied relative to ground. The voltage at the upper end is v_{in}. Since the output voltage of the scalor amplifier is -1 times its input voltage, the voltage at the lower end of the resistance voltage divider must be $-v_{in}$ relative to ground. We may use this known relation to simplify the circuit somewhat, as in Figure 4.16.

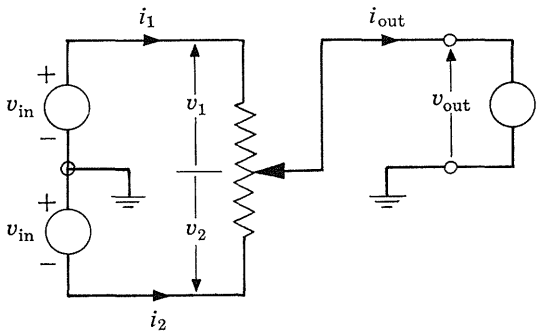
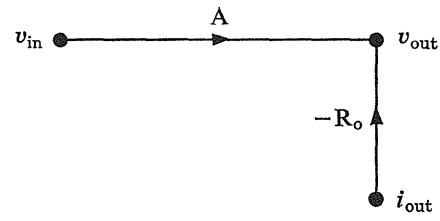


FIGURE 4.16

We have shown that by properly scaling the signals to be "multiplied," it should be possible to use an adjustable resistance to perform this operation provided some means

can be found to move the sliding contact so that $2k - 1$ is proportional to the value of the signal $g(t)$ at every instant. In practice, one uses a small servomotor which turns in one direction or the other depending on the polarity of the voltage applied to the motor. By making this voltage depend on the discrepancy between $g(t)$ and $b[2k(t) - 1]$, it is possible to cause the contact to move so as to reduce this to zero. Then, $b[2k(t) - 1] = g(t)$, and the value of the *open-circuit* output voltage is proportional to the product of $f(t)$ and $g(t)$.

QUESTION 4.16 Construct a flow-graph which expresses i_1, i_2, v_1, v_2, and v_{out} in terms of the sources v_{in} and i_{out}. Reduce this graph so as to express the open circuit voltage ratio A and equivalent output resistance R_o as functions of the parameters k and R. (Answer)



QUESTION 4.17 To obtain the greatest precision when using analog computers, it is good practice to use the largest signals possible without producing an overload condition. For instance, the maximum signal voltage in the TR-10 should not exceed 10 V in magnitude. Likewise, the value of k in the multiplier must lie between 0 and 1. Let a, b, and c be three scaling factors defined by the relations

$$f(t) = av_{in}(t),$$

$$g(t) = b[2k(t) - 1],$$

$$r(t) = cv_{out}(t).$$

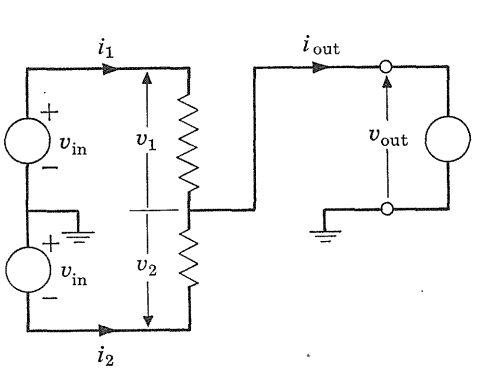
For given functions $f(t)$ and $g(t)$, the maximum values of the computer signals v_{in}, v_{out} and k will depend on the choice of the scaling constants a, b, and c.

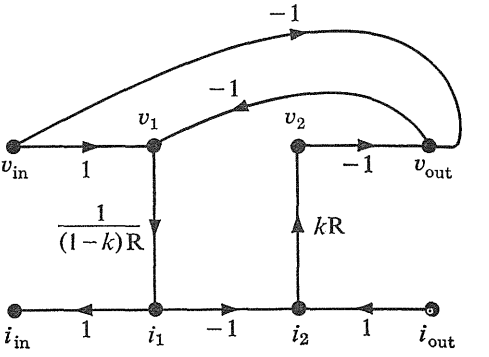
Suppose that the maximum absolute value of $f(t)$ is 50 and the maximum absolute value of $g(t)$ is 0.2.

1. What values would you select for the scale factors a and b to achieve the greatest possible precision?

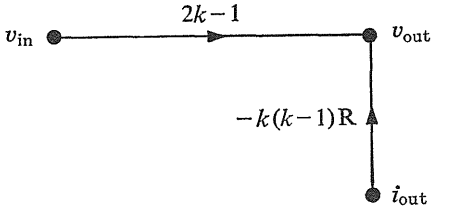
2. To find the value of the product $r(t) = g(t)f(t)$, you must multiply the value of the output voltage $v_{out}(t)$ by a scale factor c. What is its value for the choice of scale factors a and b considered in part 1? (Answer)

ANSWER TO QUESTION 4.16 By application of Kirchhoff's laws, (for example, $v_1 = v_{in} - v_{out}$, $v_{out} = -v_{in} - v_2$, etc.), construct the graph





On reduction, we obtain



$$A = 2k - 1,$$
$$R_o = k(k-1)R.$$

Here, the open-circuit voltage ratio A may be varied between -1 and $+1$ by changing k from 0 to 1. When k is 0.5, the output voltage at the midpoint is 0 because the effects of the opposing voltages applied at the opposite ends of the potentiometer exactly cancel.

ANSWER TO QUESTION 4.17

1. Since the largest value of v_{in} is 10, we may choose a to make $f(t) = 50$ correspond to $v_{in} = 10$:

$$f(t) = av_{in}(t),$$
$$50 = a\,(10),$$
$$\therefore \mathbf{a = 5.}$$

Similarly, for $k(t)$, the maximum absolute value of $g(t)$ is 0.2. Hence,

$$g(t) = b[2k(t) - 1],$$
$$\therefore \mathbf{0.2 = b.}$$

2. The scale factor c is obtained from the equality

$$\begin{aligned} r(t) &= f(t)g(t) \\ &= [av_{in}(t)][b(2k(t) - 1)] \\ &= ab[v_{in}(t)(2k(t) - 1)] \\ &= abv_{out}(t) \qquad \therefore c = ab \\ &= \mathbf{(5)(0.2) = 1.} \end{aligned}$$

A disadvantage of this simple scheme is that the voltage divider will deliver the correct output voltage only to an *open circuit*. A load current i_{out} will introduce an error in the output voltage by an amount proportional to a nonlinear function, $k(1 - k)$, of k. To circumvent this difficulty, we may mechanically fasten together two or more identical potentiometers so that the k-values are the same for each of them. Then, if the same load resistance is connected to each moving contact, the same *proportional* error will be introduced in each voltage divider. But if the fractional error is the same for each circuit, we may compensate for this error by changing k slightly. All of this is accomplished automatically by the *servomultiplier* arrangement illustrated in Figure 4.17.

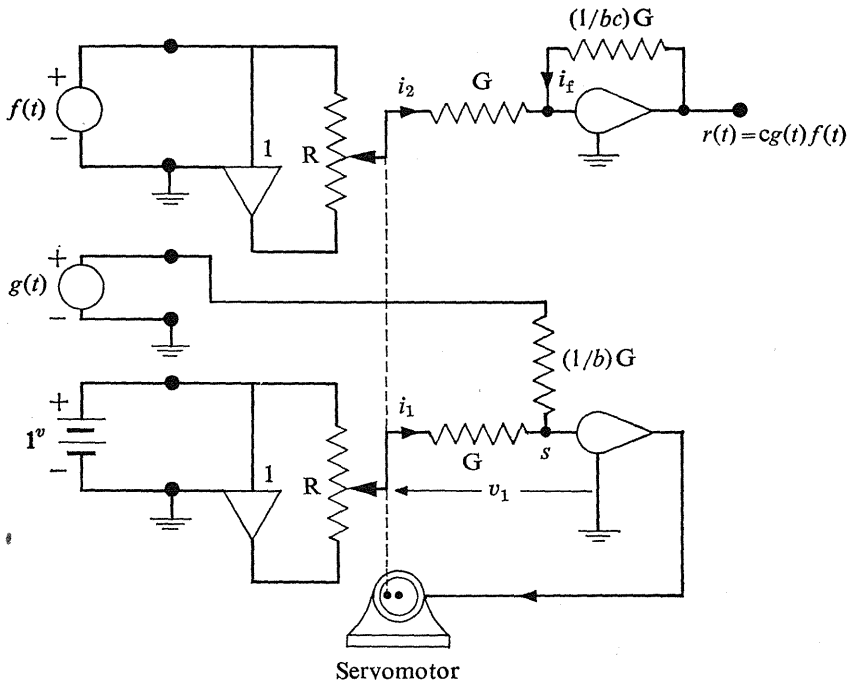


FIGURE 4.17 *A servomultiplier.*

In the arrangement of Figure 4.17, the lower potentiometer is part of a servomechanism which adjusts the moving contact so that the (output voltage)/(input voltage) ratio of *both* potentiometers is proportional to $g(t)$. The load current i_1 through the summing conductance of the lower operational amplifier is proportional to the voltage v_1 to ground of the lower sliding contact. This current and voltage is related to k by the flow-graph previously derived, shown in Figure 4.18. If the *total input* current into the input junction s of an operational amplifier differs from zero by even the slightest amount, the amplifier will deliver a very large output which in the arrangement of Figure 4.17 will cause the servomotor to move the sliding contact. This movement will be in one direction if the total current is positive, and in the opposite direction if the current is negative. These directions may be arranged so that the movement will tend to reduce the total

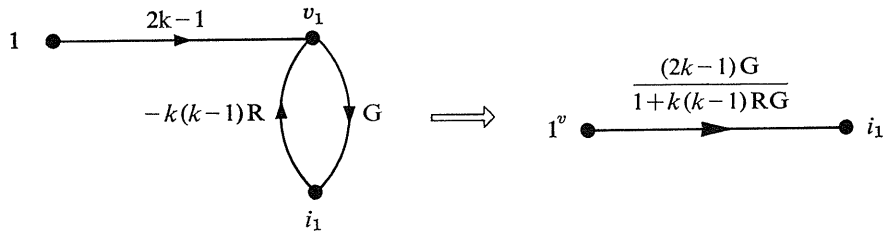


FIGURE 4.18

input current to zero. When the input current reaches zero, the output of the amplifier likewise becomes zero and, ideally, the movement ceases.*

The *total* input current into the summing junction of the lower amplifier is composed of not only the current i_1 from the potentiometer but also a current which is proportional to the voltage g. Hence, the total current into the summing junction is

$$i_1 + (G/b)g = i_s.$$

But in proper operation, the servomotor will have modified the current i_1 so as to make the *total* current i_s zero. Hence, at any instant

$$i_1(t) = -(G/b)g(t).$$

In other words, the relation between the unit input signal and the current i_1 is expressed by a transmittance of value $-(G/b)g(t)$, which is directly proportional to the value of the signal g at any instant, as shown in Figure 4.19.

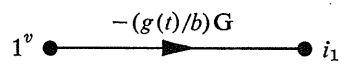


FIGURE 4.19

The top potentiometer is identical to the bottom one just considered. The transmittance which relates the current i_2 into the summing junction of the upper operational amplifier must also vary with $g(t)$ in the same manner as that of the lower amplifier. Hence, the value of i_2 at any instant is expressed by the relation illustrated by Figure 4.20.

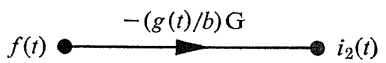


FIGURE 4.20

Here, again, the output voltage of the upper amplifier will "adjust" itself so that the current i_f fed back will be equal and opposite to i_2. If we denote the output voltage of the upper amplifier by $r(t)$, the zero-summing-current condition will be achieved when

$$i_f = -i_2,$$
$$i_f(t) = (G/b)g(t)f(t).$$

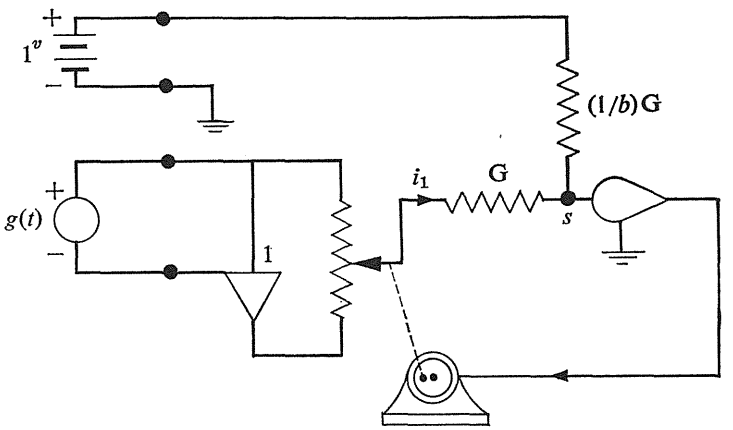
* Actually, the mechanical inertia of the motor and gears may cause the motion to "overshoot" the *null* point, whereupon the operational amplifier will drive the motor in the reverse direction, back toward the null position. However, overshoot may again occur, this time in the opposite direction, and indeed the slider may continue to "hunt" indefinitely around the true null position. We shall consider ways of eliminating this difficulty in a later chapter.

But, $i_f = (G/bc)r$. Hence,

$$r(t) = cg(t)f(t). \tag{4.4}$$

In this way, we may build an adjustable voltage divider which adjusts its transmittance to be proportional to a signal $g(t)$. Evidently, although the value of the signal g must vary slowly with time for the servomotor to follow accurately, the value of the other signal f may change much more rapidly, being limited only by the response characteristic of the operational amplifiers.

QUESTION 4.18 Suppose that in the servomultiplier just described, we interchange the unit signal with the signal g which feeds the lower potentiometer arrangement.



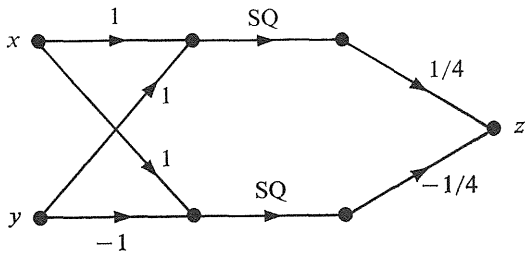
What will be the value $r(t)$ of the output signal of the upper operational amplifier (not shown), when the input signal to the upper divider has a value $f(t)$? (Answer)

Quarter-square multiplier • The servomultiplier just described may be made very accurate when $g(t)$ is a slowly varying function of time. However, if $g(t)$ changes so rapidly that the servomotor cannot move the sliding contacts to the null position, then the transmittance will not be exactly proportional to $g(t)$ and the output signal will be in error. Purely electrical systems, which do not have moving mechanical parts, are inherently capable of responding much more rapidly than those that do have them. Whereas the response time for mechanical systems is typically described in seconds, the response time of purely electrical systems is commonly of a duration of a few millionths of a second, or less. Hence, purely electronic multipliers should be capable of "multiplying" two signals whose values are changing very rapidly.

A basic method for constructing an electronic multiplier uses the *nonlinear* relationship between the input and output signals exhibited by some electrical devices. We have spoken earlier of the squaring operator SQ. This may be realized electrically in a variety of ways. The conductance of some materials (such as thyrite) increases directly with the magnitude of the voltages so as to yield a current-voltage relation that is nearly parabolic in shape, that is, $I = k|V|V$. Also, special vacuum tubes have been designed in which an electron beam passes through a mask with parabolic-shaped holes, thus yielding an

output current which is a quadratic function of the input voltage (which deflects the beam across the mask).* These special vacuum tubes have been used successfully to "square" *input* signals having frequencies as high as 40 Mc/sec and a dynamic range of 50 dB.

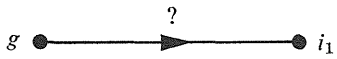
QUESTION 4.19 Consider the following operator graph. It contains scalors, with the transmittances shown, and two squaring operators. What is the value of the output signal $z(t)$ at any instant t, when the values of the input signals at that instant are $x(t)$ and $y(t)$? (Answer)



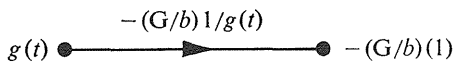
ANSWER TO QUESTION 4.18 The servomotor can come to rest only when the total current into the summing junction of the operational amplifier is zero. This requires that

$$i_1 = -(G/b)(1).$$

The equivalent transmittance of the potentiometer arrangement



must be equal to the value of the output $i_1(t)$ divided by the value of the input:



Then, by precisely the same reasoning previously employed, the value of the output signal r will be

$$r(t) = c\frac{f(t)}{g(t)}.$$

Hence, by simply interchanging these two signals, a servomultiplier may also function as a *divider*.

More generally, suppose that we replace the unit source by a voltage of value $h(t)$. The output voltage will then have a value given by

$$r(t) = c\frac{f(t)h(t)}{g(t)}. \tag{4.5}$$

* A. S. Soltes, "A Wide-Band Square-Law Circuit Element," *IRE Trans.* ED-2 No. 2 (April 1955), pp. 32–39. Also, "A Wide-Band Square-Law Computing Amplifier," *IRE Trans.* EC-3, No. 2 (June 1954), pp. 37–41.

Hence, a single servomultiplier can simultaneously "multiply" and "divide" by the values of $h(t)$ and $g(t)$, respectively. Furthermore, several potentiometer arrangements may be connected together, thus permitting several different $f(t)$'s to be simultaneously multiplied by the same factor.

One-quadrant multiplier using LOG and EXP operators • Another scheme for "multiplying" two signals makes use of the LOG and EXP operators, as shown in Figure 4.21. This realization is less useful than the quarter-square method because the values of

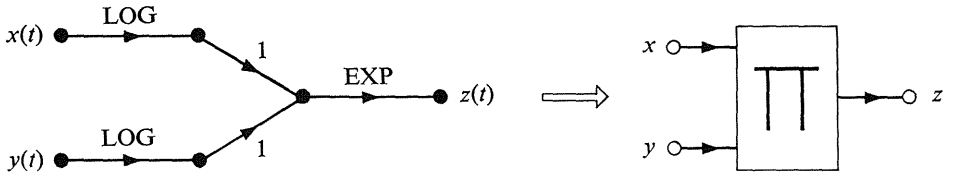
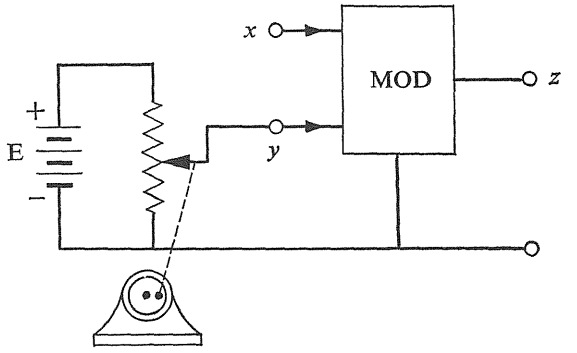


FIGURE 4.21 *Another realization of a multiplier.*

both input signals to the LOG operators must be positive at all times. This arrangement is sometimes called a "first-quadrant" multiplier because the allowable values of the two input signals correspond to the coordinates of a point lying in the *first* quadrant of the xy-plane (where $x > 0$ and $y > 0$).

QUESTION 4.20 Consider the following arrangement in which the value of y changes with time



$$y(t) = k(t)E,$$

where

$$k(t) = \frac{1}{2}\left[1 - \cos\left(\frac{\pi t}{30}\right)\right].$$

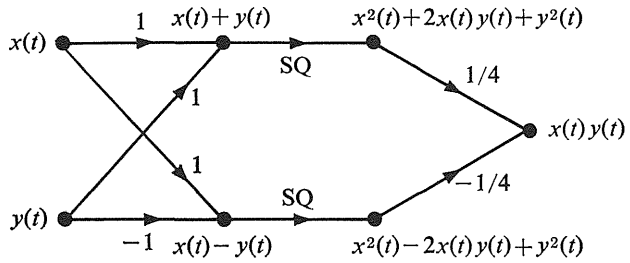
The signal z evidently depends on the first input signal x. What is the value of z at any instant t in terms of the value $x(t)$ of the first input signal? May z be represented by a *linear* operator on x? (Answer)

How may a LOG or EXP operator be realized physically? One possibility makes use of the fact that for a junction of *semiconductor* materials (such as that used in transistors), the relation between the values of the current I and voltage V is ideally given by the equation

$$I = I_0[e^{kV} - 1], \tag{4.7}$$

where k and I_0 are constants which depend on the physical properties and configuration of the semiconductor material. Hence, we may describe the current–voltage relation by the *static, nonlinear* graph illustrated in Figure 4.22.

ANSWER TO QUESTION 4.19 The operator graph shows one of the main methods of producing a physical multiplier or *modulator* operator. The SQ operator is static, so we need know only the value of the input signal at the instant t to

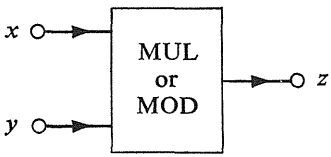


find the output at that same instant. The values of all the signals are shown in the accompanying flow-graph at their respective nodes. Hence, at any instant t the value of the output signal z of a modulator is equal to the product of the values of the two input signals x and y:

$$x(t)y(t) = z(t). \tag{4.6}$$

The physical modulation process thus corresponds to the arithmetic multiplication. (In fact, in most analog computers, modulators are actually called *multipliers*. However, to preserve the different meanings of *signal* and signal *value*, we shall refer to the physical process as a *modulator*.)

A modulator (or "multiplier") is a different kind of operator than those we have considered hitherto—it operates on *two* signals to yield *one* signal (for this reason it is sometimes called a *binary* operator). Hence, our regular branch notation breaks down and we must adopt a symbol having two inputs and one output such as the one illustrated:



where $x(t)y(t) = z(t)$.

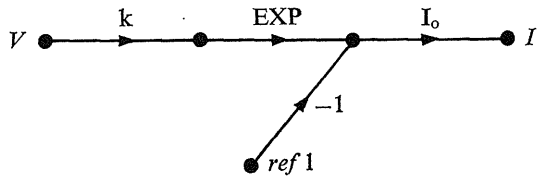
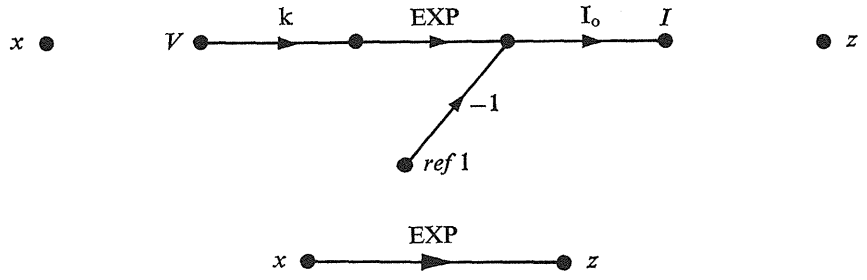


FIGURE 4.22 *A semiconductor diode operator, where the reference signal ref* 1 *has a constant value of ref* $1(t) = 1$ *for all* t.

QUESTION 4.21 Plot a graph of I as a function of V for $k = 1$ and $I_0 = 1$ over the interval $-3 \leq V \leq 3$. (Answer)

QUESTION 4.22 We wish to realize the EXP operator, starting with the semiconductor operator shown in Figure 4.22. Determine the scalor branches that may be incorporated between the external signal nodes x and z and the V and I nodes of the semiconductor operator shown herewith to yield the exponential operator:



The LOG operator may be realized by inverting EXP using the method already described, as shown in Figure 4.23.

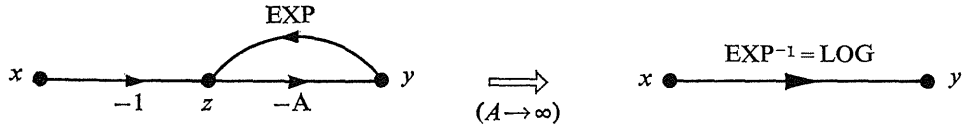


FIGURE 4.23 *Realization of the* LOG *operator using an* EXP *operator.*

As A is made infinitely large, the value of z must approach zero at every instant. This implies that

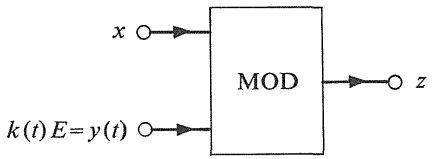
$$z = -x + y\text{EXP} = 0,$$

$$y\text{EXP} = x,$$

or

$$y = x\text{EXP}^{-1} = x\text{LOG}. \tag{4.8}$$

ANSWER TO QUESTION 4.20 If the relation between x and z is given by a linear operator, it must exhibit *additive* and *homogeneous* properties. From the basic definition of a modulator, the value of the output signal is given by



$$x(t)y(t) = z(t),$$
$$x(t)k(t)E = z(t).$$

To show additivity, suppose that

$$x_1 + x_2 = x$$

or

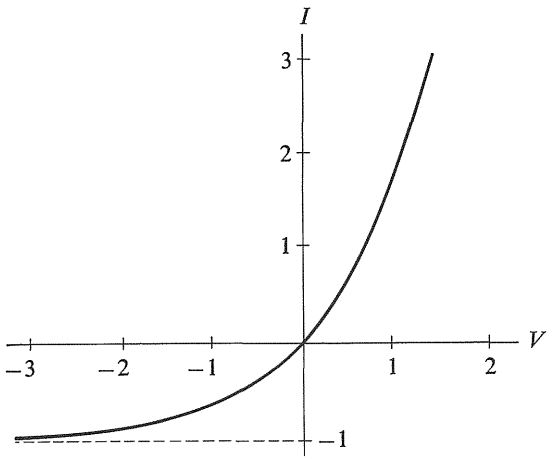
$$x_1(t) + x_2(t) = x(t).$$

Then,

$$x_1(t)k(t)E + x_2(t)k(t)E = z(t).$$

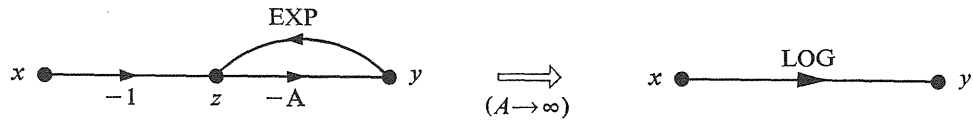
But the two terms on the left are the output values that would have been obtained with only x_1, or only x_2, respectively; so additivity is thus established. Similarly, if the input signal were ax instead of x, it is obvious that the output signal will be az instead of z. Consequently, the modulator is a linear operator (also static and *non*stationary).

ANSWER TO QUESTION 4.21 The plot of the function $I = e^V - 1$ is as shown. Clearly, the relation is highly nonlinear. Except for the subtractive constant,



I_0, the relation between I and V is exponential. It also shows that the semiconductor diode conducts strongly in one direction and weakly in the other.

QUESTION 4.23 In the following realization of the LOG operator shown below, why must the value of x always be positive?



(Hint: Remember that the value of y must always be finite. Since $z = -yA^{-1}$ it is clear that as $A \to \infty$, the value of z must become vanishingly small.) (Answer)

QUESTION 4.24 Using only scalors and EXP operators, design a "divider" which for input-signal values $x(t) > 0$ and $y(t) > 0$, will deliver an output signal whose value is

$$r(t) = \frac{x(t)}{y(t)}.$$

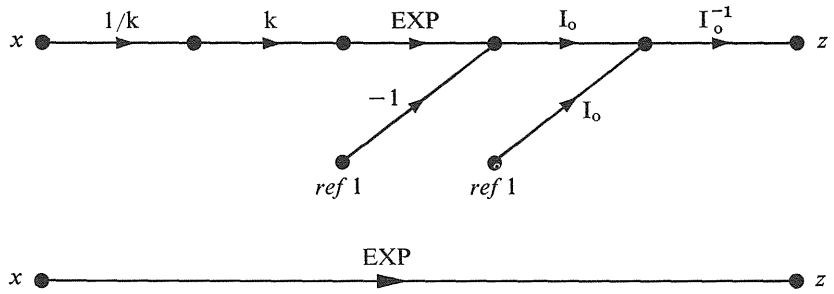
(Answer)

QUESTION 4.25 Using only scalors and SQ and EXP operators, design a system with inputs $x(t)$ and $y(t)$ (where $y(t) > 0$), which will yield an output signal of value

$$r(t) = \frac{y(t)}{\sqrt{x^2(t) + y^2(t)}}. \tag{4.9}$$

(Answer)

ANSWER TO QUESTION 4.22 Three scalor branches must be incorporated to realize the operator EXP from a semiconductor diode. One of these is used to cancel out the "saturation" current I_0; the other two scalors compensate for the input

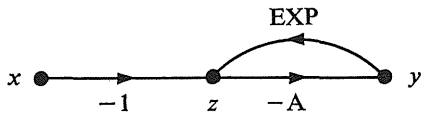


and output scale factors, as shown. One of the difficulties with the arrangement is that exact cancellation of I_0 in the output may be difficult to achieve and maintain. (This is a common problem in the design of electronic circuits.)

Three Basic Operators (Scalors, Limitors, Delayors)

The preceding questions have illustrated how we may start with a set of given operators and suitably arrange them in a system to construct new kinds of operators. This is an example of an important fundamental idea in studying systems of all kinds. One starts with a limited variety of building blocks, and then arranges them into more complex structures. Each type of building block has its own peculiar properties which we must be able to represent and describe in terms of relations among the observables. But also, we must be able to represent and describe the arrangement of these building blocks into larger systems. This must be done when representing any system. For instance, in chemistry the atomic elements might be considered to form a finite set of building blocks, each of which has been denoted by a symbol and represented in terms of its observable characteristics such as its valence, atomic weight, etc. These elements may then be connected together to form "new" compounds which exhibit properties that may be much more complex than those of the individual elements. Likewise, in programming a digital computer, there are only a limited set of basic operations available that the machine can perform, and somehow these must be used repeatedly and in the right sequence to perform much more complex operations and to achieve the desired result.

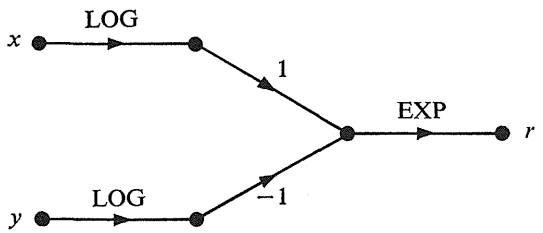
ANSWER TO QUESTION 4.23 The value of the signal at node z approaches zero as $A \to \infty$. This follows from the fact that y is finite and $-z\mathrm{A} = y$. Also,



$$-x + y\mathrm{EXP} = z.$$

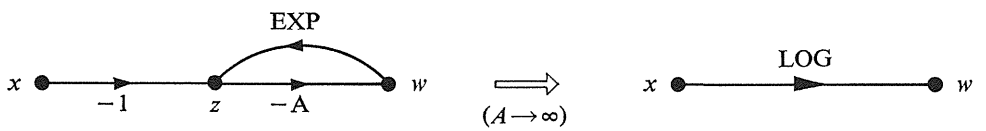
The value of $y\mathrm{EXP}$ is never negative, $y\mathrm{EXP}(t) = \mathrm{e}^{y(t)}$. For $z(t)$ to be zero, it is necessary that $-x(t)$ be negative, or that the value of the input signal x be positive at every instant.

ANSWER TO QUESTION 4.24 A possible design for a system that will exhibit an output-input relation of the form $r(t) = x(t)/y(t)$ is



$$x(t) > 0$$
$$y(t) > 0$$

where each LOG operator is realized from EXP operators by

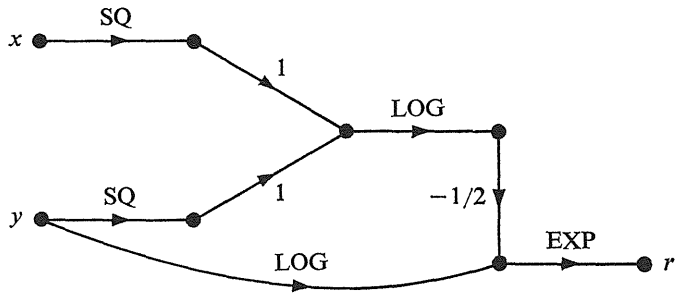


We have introduced signal flow-graphs as a useful symbolic scheme for describing systems of operators, and we have illustrated various kinds of linear and nonlinear operators. Since it is possible to start with any one of several different sets of basic building blocks to concoct systems which will be equivalent to other kinds of building blocks one might think that the choice of basic operators is completely arbitrary. Although the choice is by no means unique, certain operators may rightfully be considered as more "basic" than others because of their logical simplicity, generality, and ease of physical realization.

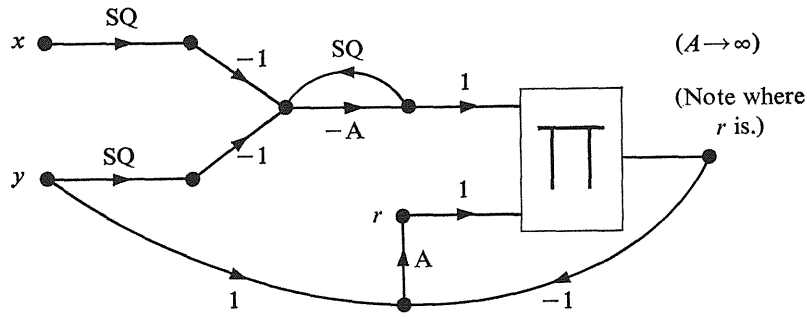
Certainly, it seems proper that one of the basic kinds of operators should be the scalor. The scalor yields an output signal whose value at any instant is proportional to the value of the input signal at the *same* instant.

To provide a representation for the influence of the past on the present, we need another basic kind of operator, the *delayor*, which yields an output signal whose value at any instant is equal to the value of the input signal at an earlier instant (of constant age).

ANSWER TO QUESTION 4.25 A possible design for obtaining the output-input relation $r(t) = y(t)/(x^2(t) + y^2(t))^{1/2}$ is

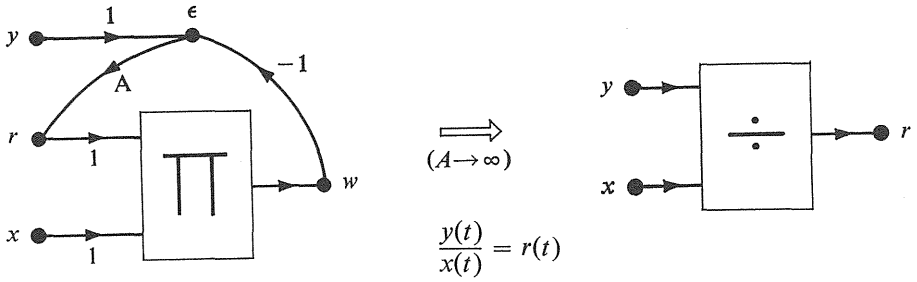


where LOG is realized from EXP by the inversion method. There are many possible configurations. For instance, it is possible to use only scalors and SQ operators, such as



where the "multiplier" is realized using SQ-operators as already described. In this arrangement, we have realized a divider by "inverting" one input to a "multiplier."

The arrangement is shown herewith. For very large positive values of the transmittance



$$\frac{y(t)}{x(t)} = r(t)$$

A, the value of the signal at ϵ must be essentially zero at all times. This requires that the output w of the multiplier must be essentially equal to the *input* signal y. Therefore, at any instant

$$x(t)r(t) = w(t) = y(t)$$

or

$$r(t) = \frac{y(t)}{x(t)}. \tag{4.10}$$

In order that the small error signal ϵ will affect w so as to *reduce* ϵ to zero, it is necessary that $x(t)$ be positive. Hence, unless some means is found for changing the algebraic sign of A when $x(t)$ becomes negative, this divider will operate correctly only for positive-value input signals.

By combining a scalor and a delayor, we obtain an operator which yields an output value that is proportional to the value of the input of a *prior* instant. We shall show in a subsequent section that by using *only* scalors and delayors, we may describe the enormous class of linear, stationary operators with which you will be largely concerned.

Whereas the choice of scalors and delayors as basic linear operators is relatively obvious, the best choice of basic *nonlinear* operators is much less certain. In this book, we use the static *limitor* as the basic nonlinear operator. By using limitors in conjunction with scalors, we may approximate as closely as we wish most nonlinear operators (such as SQ and EXP). We select the limitor because it corresponds to a significant nonlinear property exhibited by many physical devices.

Limitors

We define a limitor to be a static, stationary operator which has unit transmittance when the input signal value is positive and has zero transmittance when the input value is negative. To denote a limitor, we shall use either of the symbols in Figure 4.24. It is

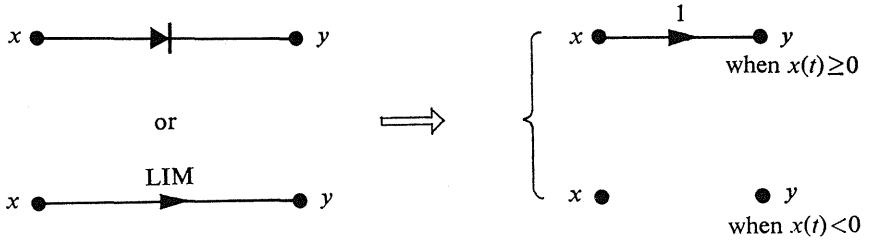
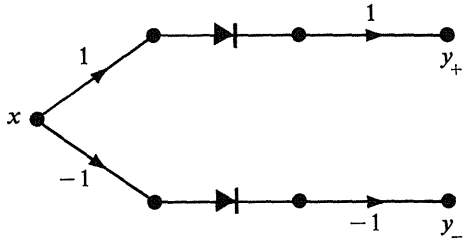


FIGURE 4.24 *Definition of a limitor.*

evident that a limitor transmits only that portion of a signal which is positive in value, and blocks completely the negative part of the input signal. Thus, we can separate a signal into its positive and negative portions by implementing the operations illustrated by Figure 4.25.

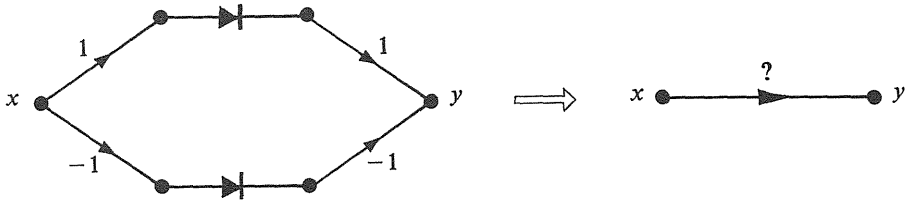


The value of y_+ is equal to $x(t)$ when $x(t) \geq 0$. Otherwise $y_+(t)$ is zero.

The value of y_- is equal to $x(t)$ when $x(t) \leq 0$. Otherwise $y_-(t)$ is zero.

FIGURE 4.25

QUESTION 4.26 Is the operator shown below equivalent to a simple operator? If so, what is it? (Answer)



QUESTION 4.27 Using only limitors and scalors, can you construct the absolute-value operator ABS previously encountered in Question 4.12? (Answer)

The limitor enables us to separate a signal into its positive-valued and negative-valued components. This makes it possible to extend the operating range of the multiplier using LOG and EXP operators to two and four quadrants. For instance, to provide two-quadrant multiplication for either positive and negative values of $x(t)$, with $y(t) > 0$,

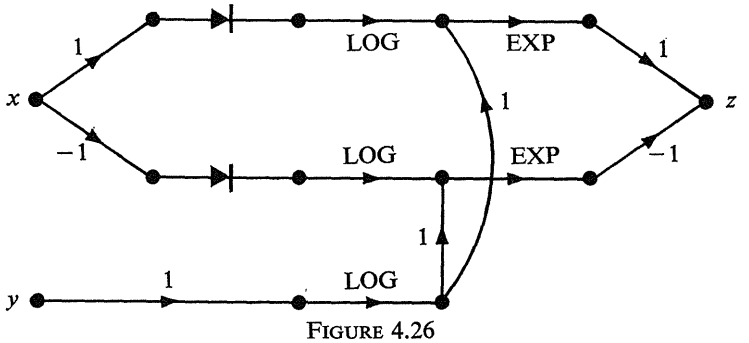


FIGURE 4.26

we may expand the operator considered earlier as illustrated by Figure 4.26. You should be able to show without any difficulty that

$$z(t) = x(t)y(t) \qquad \text{for } y(t) > 0.$$

QUESTION 4.28 Modify the design of the two-quadrant multiplier just discussed to permit four-quadrant operation with any x and y.

Approximation of static nonlinear operators using limitors and scalors • If you plot pairs of input-output values for a limitor, you will obtain the graph shown in Figure 4.27. This operator "clips off" all input-signal values less than zero; that is, whenever the

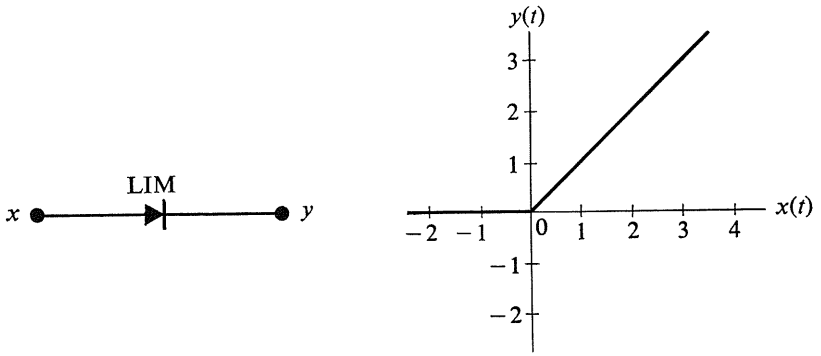


FIGURE 4.27 *Input-output characteristic of a limitor.*

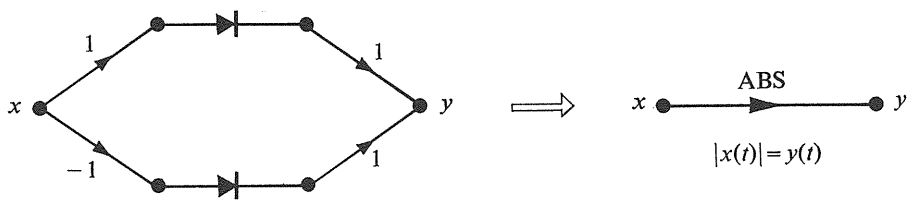
input signal has a negative value, the output signal value is zero. Sometimes it is useful to clip off the input signal below some threshold level, z, other than zero. This is readily accomplished by the arrangement in Figure 4.28. The clip level z may be a constant such as 1, or it may be determined by the value of another signal. In either case, the above arrangement defines a new operator CLIPz which yields at its output a signal

ANSWER TO QUESTION 4.26 The value of the output signal y is

$$y_+ + y_- = y.$$

When $x(t)$ is positive, $y_+(t) = x(t)$ and $y_-(t) = 0$. When $x(t)$ is negative, $y_-(t) = x(t)$ and $y_+(t) = 0$. Clearly, $y(t)$ is identical to $x(t)$ for all positive and negative values. (Furthermore, $y(t) = 0$ when $x(t) = 0$.) Hence, when combined, these two *nonlinear* limitors are equivalent to a *linear scalor*. This possibility of combining two nonlinear elements to obtain an overall system that is linear is widely used in electronic circuit design (e.g., in class B push–pull amplifiers).

ANSWER TO QUESTION 4.27



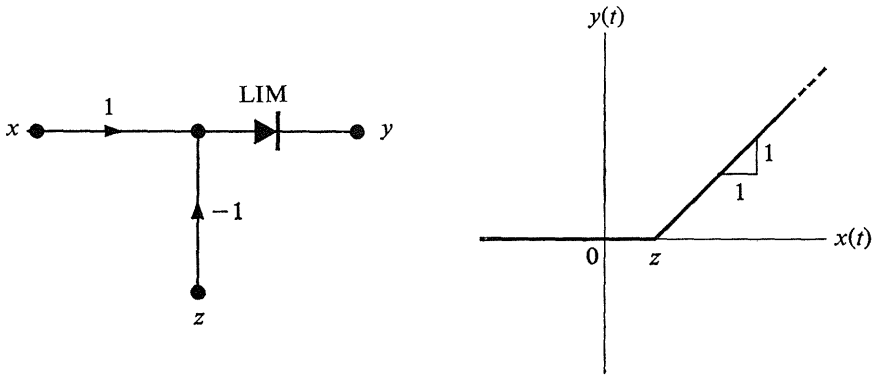


FIGURE 4.28

whose value at each instant is the amount by which the value $x(t)$ of the input signal exceeds the clipping level $z(t)$. The abbreviated symbol for this clipping operator is shown in Figure 4.29.

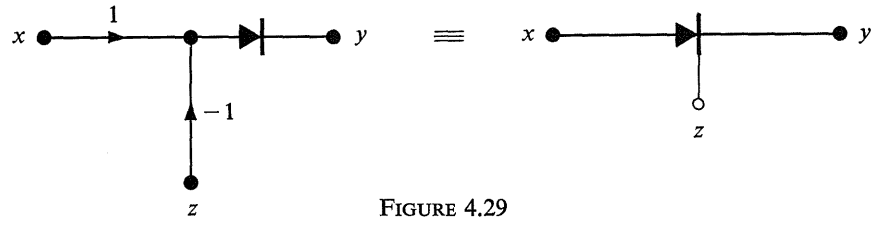


FIGURE 4.29

Or,
$$(x - z)\text{LIM} \equiv x\text{CLIP}z.$$

Observe that because LIM is nonlinear, superposition is not applicable (that is, in general, $(x - z)\text{LIM} \neq x\text{LIM} - z\text{LIM}$). Thus, at any instant, the value of the output signal is $y(t) = x\text{CLIP}z(t)$, where

$$x\text{CLIP}z(t) \equiv \begin{cases} x(t) - z(t) & \text{if } [x(t) - z(t)] > 0, \\ 0 & \text{if } [x(t) - z(t)] \leq 0. \end{cases}$$

By additively combining several clipping operators, we may approximate various nonlinear functions with straight-line segments. The shape of the segments can be made to have any desired value by placing a suitable scalor after the clipping operator.

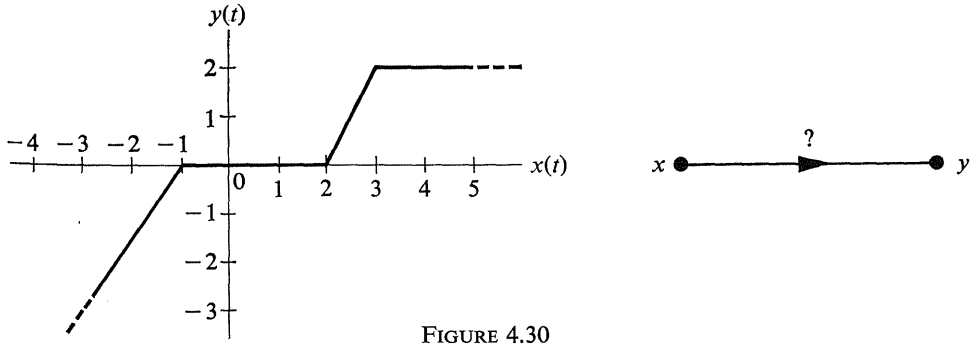


FIGURE 4.30

For instance, to realize the static nonlinear input–output characteristic plotted in Figure 4.30 we may additively combine the output of three clippers having individual input–output characteristics as in Figure 4.31.

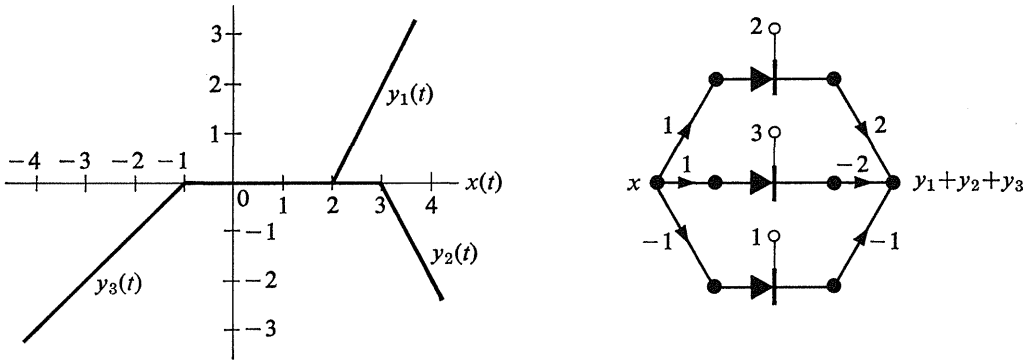
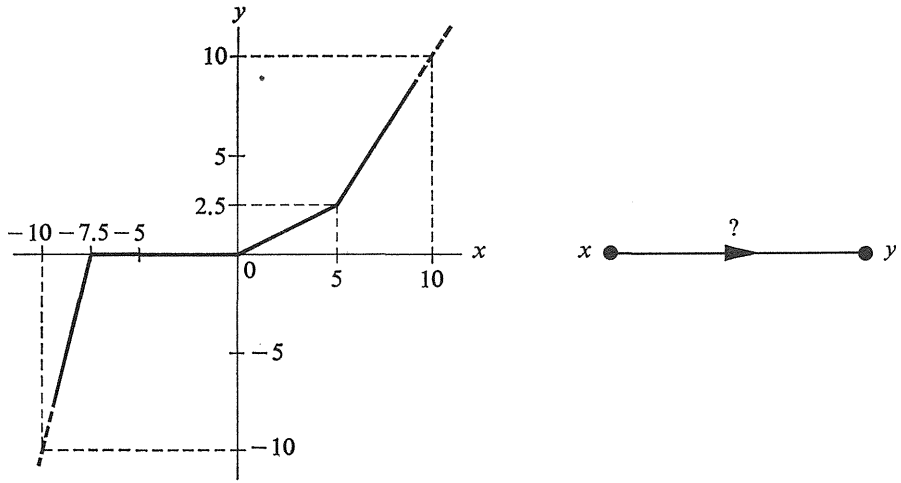


FIGURE 4.31

QUESTION 4.29 Design, using clippers and scalors, a static nonlinear operator having the input–output characteristic shown.



QUESTION 4.30 It is desired to approximate the SQ-operator for input values of x from -10 to $+10$ using an input–output characteristic composed of five straight-line segments, with breakpoints at $x = -6, -2, +2, +6$. Determine the optimum slopes of these segments so as to minimize the maximum discrepancy between $x\text{SQ}(t)$ and the approximate $y(t)$ for any $x(t)$ within the interval -10 to $+10$. What is the magnitude of this maximum error, $[y(t) - x\text{SQ}(t)]$, expressed as a percentage of the maximum value of $y(t)$? (Answer)

QUESTION 4.31 If 10 straight-line segments were used to approximate the SQ-operator considered in Question 4.30, what would be the maximum error for the best design? Can you prove that the maximum percentage error is expressed in terms of the number, n, of segments by the equation % error $= 100/n^2$?

Diode function generators • At the risk of confusing electrical circuits with signal flow-graphs, we here note that an ideal *diode circuit element* may be used to produce simple electrical circuits which exhibit some of the foregoing nonlinear input–output

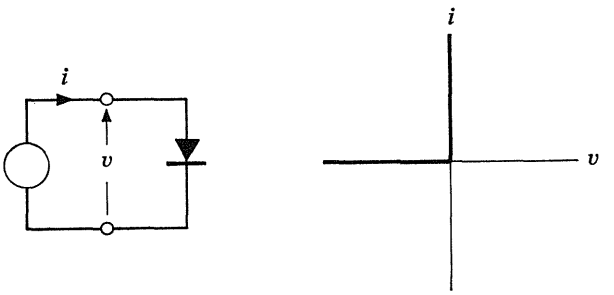


FIGURE 4.32

characteristics. An electrical diode acts like a *short* circuit when v tends to be positive and like an *open* circuit when v is negative. That is, it can carry large positive current in the direction of the arrow (•→⊢•) with very little voltage drop v (i.e., like a short circuit). On the other hand, when v is negative so that the current tends to flow against the arrow, the diode can withstand large negative voltages with very little current *flow* (i.e., like an open circuit). The trouble with the circuit of Question 4.32 is that any loading resistance connected across the output terminals will alter v_{out}. To avoid this difficulty we may introduce an operational amplifier (as in the PACE TR-10 x^2 DFG Model 16.101). As shown in Figure 4.33, the voltage at v is isolated from the summing terminals of the operational amplifier when v is negative, because the diode then acts like an open circuit.

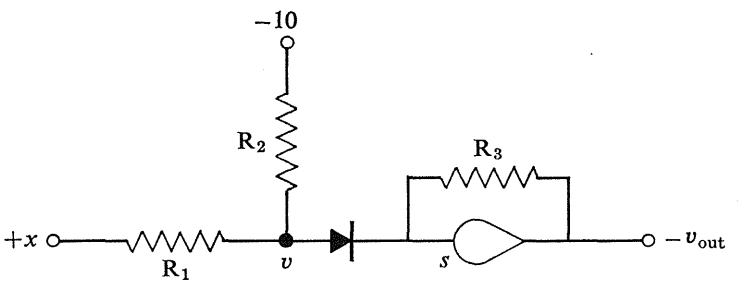
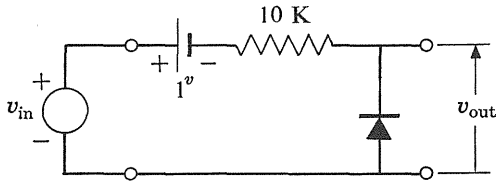


FIGURE 4.33

Under *these* conditions,

$$v = -10 + \frac{R_2}{R_1 + R_2}(x + 10) = \frac{R_2 x - R_1 10}{R_1 + R_2}.$$

QUESTION 4.32 Consider the accompanying circuit. Determine and plot v_{out} as a function of v_{in} (assuming that the diode is ideal). Note that a resistor must be



placed between the voltage source and the ideal diode to avoid a contradiction in the meaning of the symbols. Why? (Answer)

However, for $x > 10R_1/R_2$, the diode becomes in effect a short circuit and the circuit behaves like an ordinary amplifier, having an amplification of $-R_3/R_1$ for the input

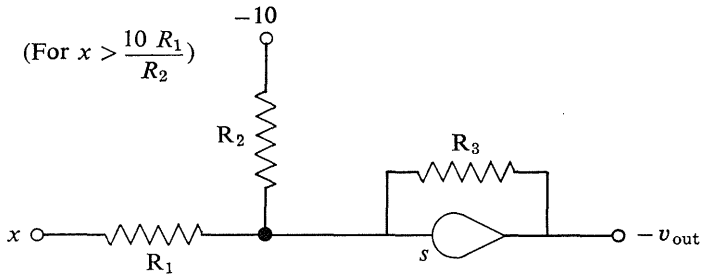
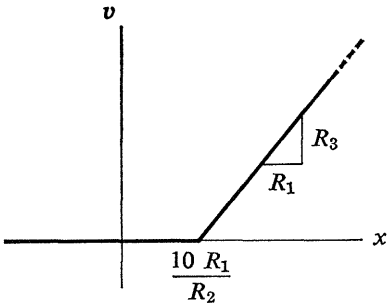


FIGURE 4.34

signal x and an amplification $-R_3/R_2$ for the negative constant -10, as in Figure 4.34. Hence, for $x > 10R_1/R_2$, the output $-v_{out}$ is

$$(-R_3/R_1)x + (-R_3/R_2)(-10) = -v_{out}.$$

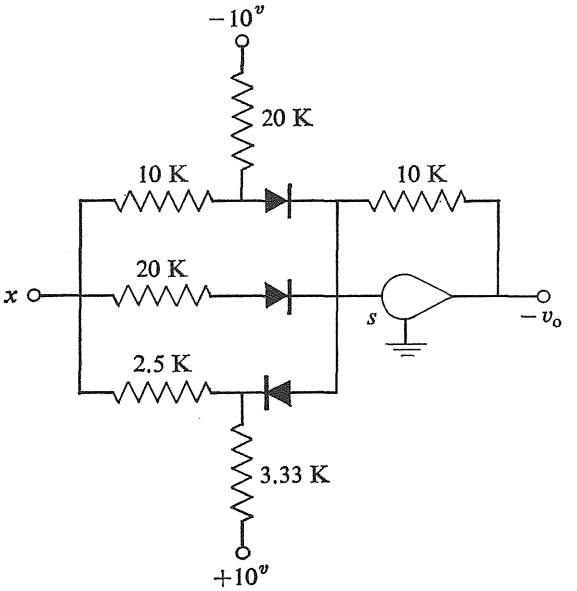
Hence,

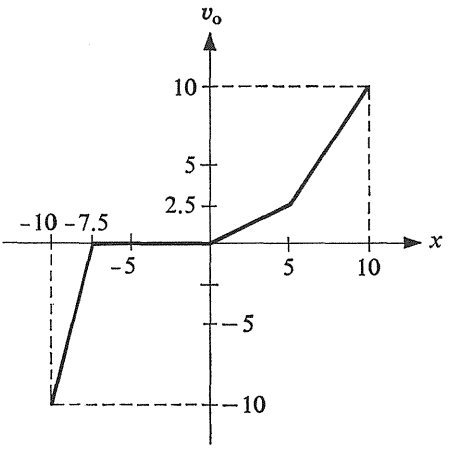


$$v_{out} = \begin{cases} (R_3/R_1)x - 10R_3/R_2, \\ 0, \quad \text{if } x < 10R_1/R_2. \end{cases}$$

FIGURE 4.35

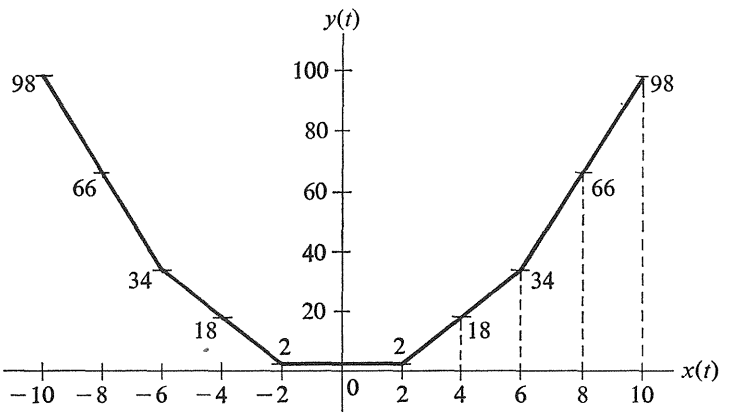
QUESTION 4.33 Show, starting from the circuit values given, that the input–output characteristic of the following circuit is as plotted. (Note that we have plotted v_o although the actual output voltage is $-v_o$, the negative of this.) (Answer)





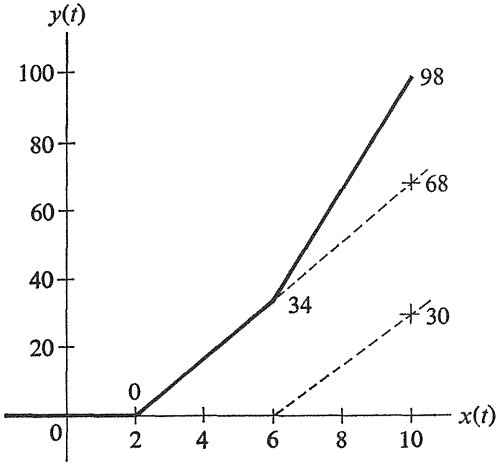
(All resistance values are in ohms)

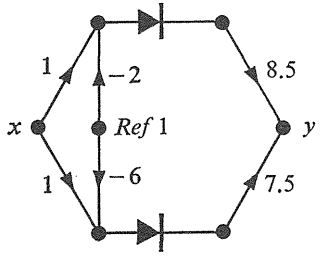
ANSWER TO QUESTION 4.30



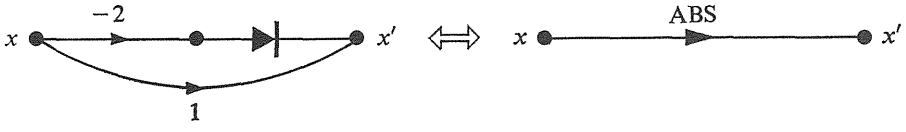
By slightly lowering each segment so that the two end points are two units *below* the exact values, the midpoint of each segment will be two units *above* the exact value. Thus, the maximum error, when minimized in this way, is only two units out of 98, or 2% of the maximum output signal.

Although this design does offer the smallest *maximum* difference between the values of the exact and approximate output signals, it has the objectionable feature of producing nonzero output of two units when the input value is zero. In practice, the central segment is lowered to coincide with the x-axis, thus increasing the maximum error to 4 when $x = 2$. (Note: *ref* 1 is a reference signal of constant, unit value.)

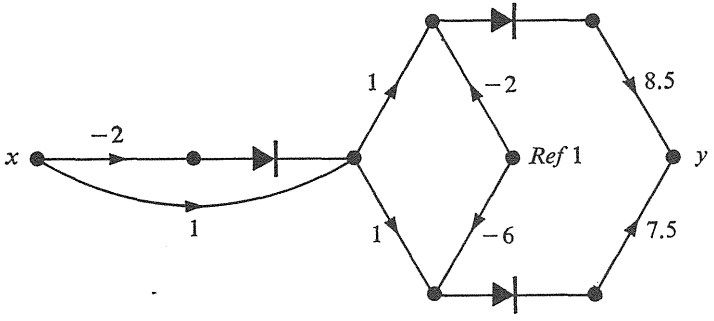




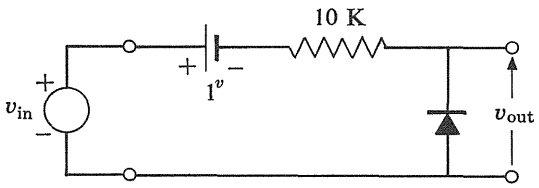
As shown above, this operator approximates SQ only for positive $x(t)$. However, since the absolute-value operator, ABS, may be realized using only one limitor by



we obtain the following simple approximation to SQ over the full range of input signal values



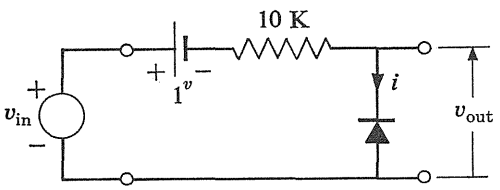
ANSWER TO QUESTION 4.32 When v_{out} is *positive*, the diode arrow is negative with respect to the diode bar, and the diode therefore acts like an *open circuit.*

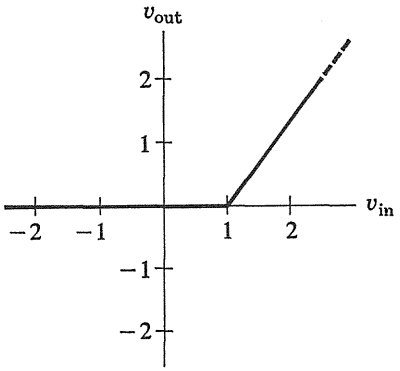


Since the diode current is then zero, the voltage across the 10K resistor is likewise zero. The output voltage therefore differs from the input voltage by the battery voltage:

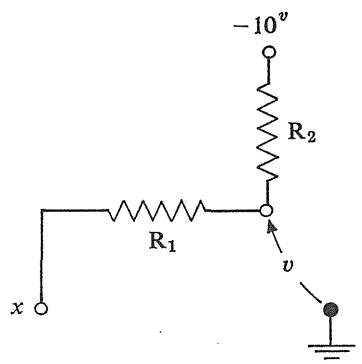
$$v_{out} = v_{in} - 1\text{ V} \qquad (\text{for } v_{out} > 0).$$

For $v_{in} < 1$ V, the output voltage would tend to become negative, but now the diode conducts current in the direction of the diode arrow and acts like a *short circuit.* The current is such that the voltage across the resistor is equal to $v_{in} - 1$ V, thus satisfying Kirchhoff's voltage law with $v_{out} = 0$.



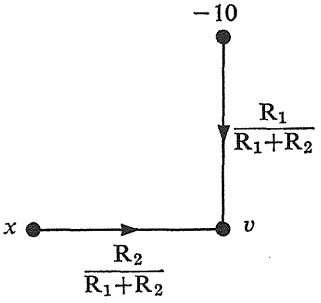


ANSWER TO QUESTION 4.33 In the circuit shown, the summing junc-



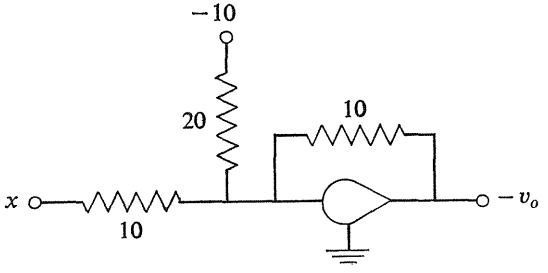
tion s of the operational amplifier is always practically at zero voltage. The diodes act either as short circuits or as open circuits, depending on whether current tends to flow through them in the direction of the diode arrow or in the reverse direction, respectively. Thus, to analyze this circuit, we must first determine whether the diode is to be treated as a short circuit or as an open circuit. Once this has been determined, the circuit is solved by the methods discussed at some length in the previous chapter.

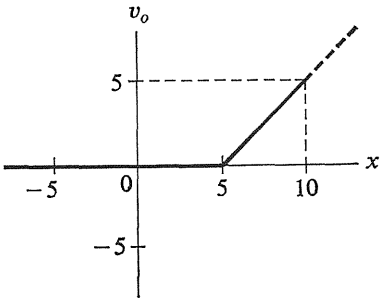
We may consider each diode branch individually. For the top branch, the voltage v is the result of the voltage x and the voltage -10. We may find v by superimposing the contribution of x (with -10 temporarily reduced to zero) and the contribution of -10 V (with x temporarily reduced to zero). The voltage-divider relation immediately yields



$$v = \frac{R_2}{R_1 + R_2}x + \frac{R_1}{R_1 + R_2}(-10)$$
$$= \frac{R_2x - R_1(10)}{R_1 + R_2}.$$

If v is negative, the upper diode will act like an open circuit, because current will try to flow against the diode arrow. For the values $R_1 = 10\text{K}$ and $R_2 = 20\text{K}$, the diode is therefore an open circuit for $x < (R_1/R_2)(10) = 5$, and there will be zero contribution to the output for $x < 5$. For $x > 5$, we have the diode acting like a perfect conductor, as shown. A similar discussion applies to the other two diode branches. Note that by reversing the electrical connections to the diode, as in the bottom branch, it may be made to conduct when x becomes more negative than -7.5.





Dead-zone operator • In most automobiles, it is necessary to turn the steering wheel through some small angle before it causes any effect on the steering. We might say that the relationship between the rotation of the steering wheel and the orientation of the front wheels is characterized by a *dead zone*. This relation is idealized in the unit dead-

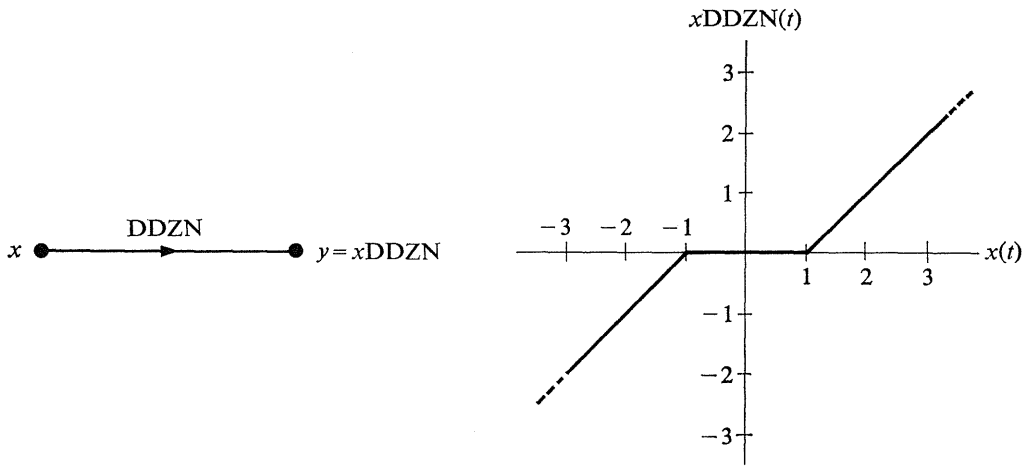


FIGURE 4.36 *Dead-zone operator.*

zone operator, DDZN, which has the input–output characteristic shown in Figure 4.36. This dead-zone operator is easily constructed using two limitors, as in Figure 4.37.

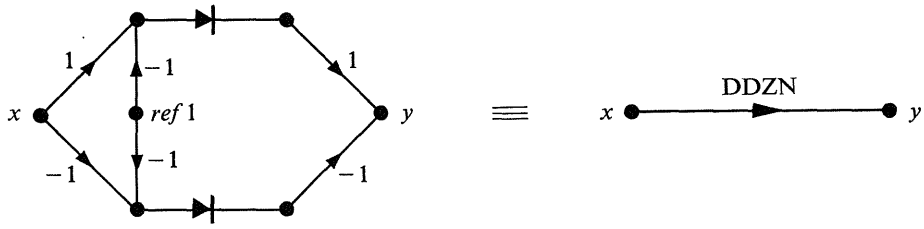
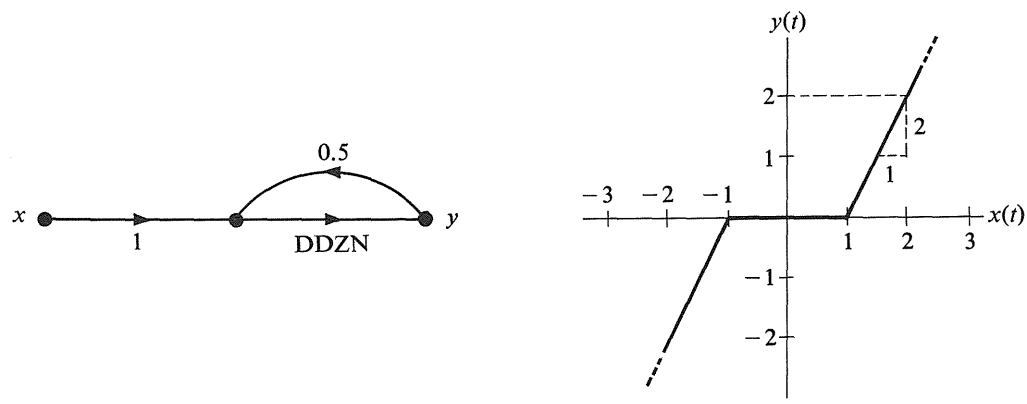
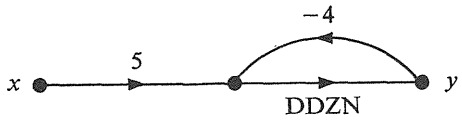


FIGURE 4.37

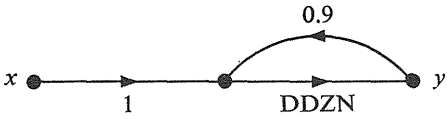
QUESTION 4.34 Show that the input–output relation of a dead-zone operator with a positive feedback transmittance of 0.5 is as shown.



QUESTION 4.35 Plot the input–output characteristic of this operator (Answer)



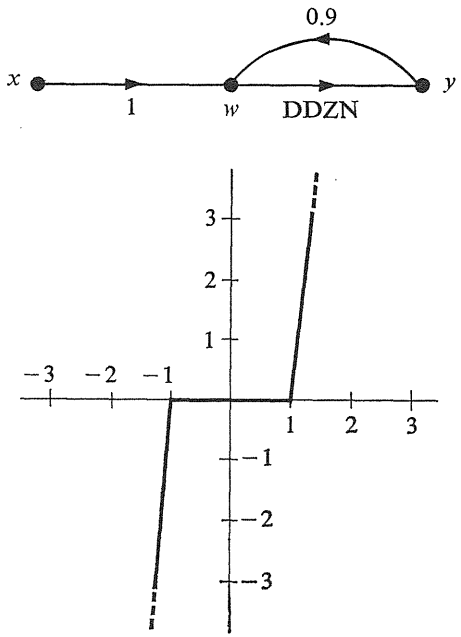
QUESTION 4.36 What is the input–output characteristic of the dead-zone operator with negative feedback around it as shown? (Hint: Assign a value to the output signal y and determine what the value of x must have been to yield this value of y.)



By comparing the answers of Questions 4.35 and 4.36, what can you conclude about the effect of negative feedback around a nonlinear operator? (Answer)

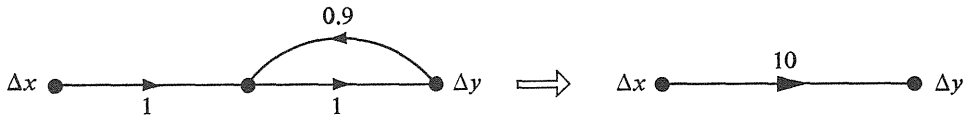
Positive and negative limitors • Many physical operators exhibit linear input–output characteristics over a limited range of input signal values, but if the input value is increased (or decreased) beyond some level, no further increase (or decrease) in the output is achieved. This situation is sometimes described by saying that the operator has become "saturated." For instance, pushing harder on a gas pedal that has already bottomed on the floorboard will have little effect on the horsepower delivered by your automobile engine. This effect is idealized by positive, negative, and symmetric limitors.

ANSWER TO QUESTION 4.35 An easy way to obtain pairs of values for x and y needed to plot the input–output characteristic is to assign a value to the output and then determine what the input must have been to produce this output.



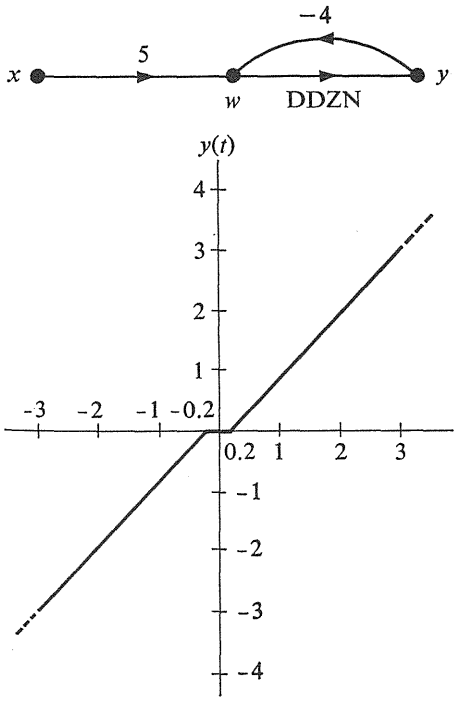
$y(t)$	$w(t)$	$x(t) = w(t) - 0.9y(t)$
5	6	1.5
3	4	1.3
1	2	1.1
0+	1	1
0	$\lvert w(t)\rvert < 1$	$\lvert x(t)\rvert < 1$
−0	−1	−1
−1	−2	−1.1
−3	−4	−1.3

In constructing the table, once the value of y has been assigned, we may use the input–output characteristic of DDZN to determine the value of w. The input x must then supply the *return difference*, $w - 0.9y$. The incremental change Δx can be estimated by considering DDZN as a scalor whose transmittance is either 1 or 0, depending upon whether the magnitude of $w(t)$ is greater than or less than 1. For $|x(t)| > 0$, we have



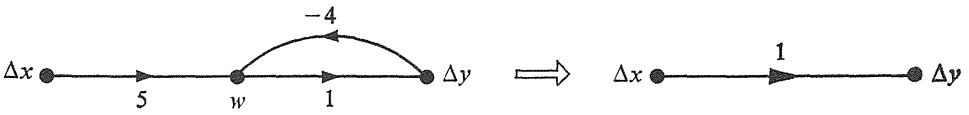
since to increase the output by 1 unit, only $\frac{1}{10}$ the required increase in the input to *DDZN* must be provided by input Δx, the rest of the input being provided by signal fed back from the output.

ANSWER TO QUESTION 4.36 Make a table as in the solution to Question 4.35. From this result, we see that the *negative feedback tends to reduce the nonlinearity of* DDZN and make it resemble more nearly a scalor having unity transmittance.

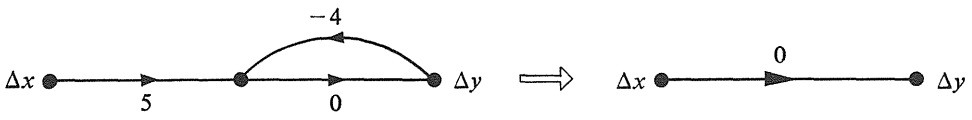


$y(t)$	$w(t)$	$x(t) = \frac{1}{5}[w(t) + 4y(t)]$
5	6	5.2
3	4	3.2
1	2	1.2
+0	1	0.2
0	$\|w(t)\| < 1$	$\|w(t)\| < 0.2$
−0	−1	−0.2
−1	−2	−1.2
−3	etc.	

The slope of this characteristic for $|w(t)| > 1$ is found by replacing DDZN by a unit transmittance, as shown.



When the value of x lies between -0.2 and $+0.2$, the incremental transmittance of DDZN is zero. Then, the slope is zero, since



A *unit positive limitor*, LMP, is defined by the input–output characteristic illustrated by Figure 4.38. More generally, we may wish to limit the output at some value z. This

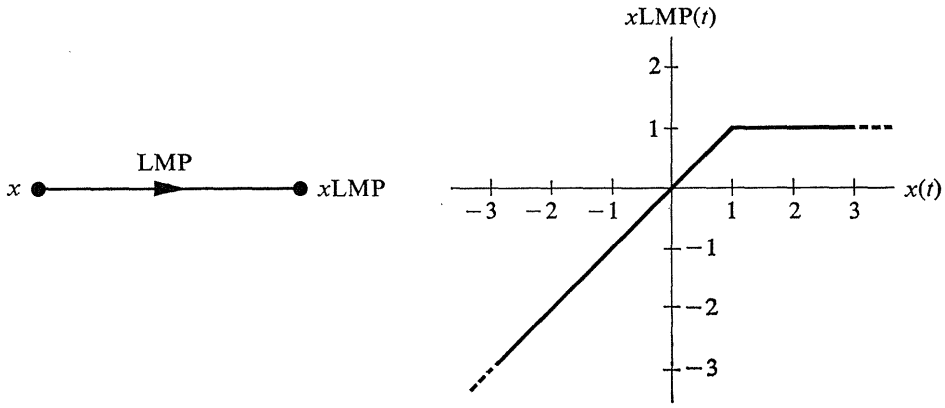


FIGURE 4.38

operator may be denoted by LMPz, where

$$x\text{LMP}z(t) = \begin{cases} x(t) & \text{if } x(t) \leq z, \\ z & \text{if } x(t) > z. \end{cases}$$

Evidently, LMP is a special case of LMPz with $z = 1$. It is easily realized as

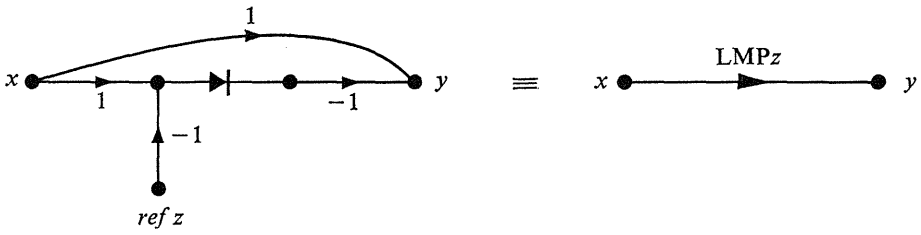


FIGURE 4.39

A *unit negative limitor*, LMN, is defined by the input–output characteristic as illustrated by Figure 4.40. More generally, the operator LMNz limits the output value to not less

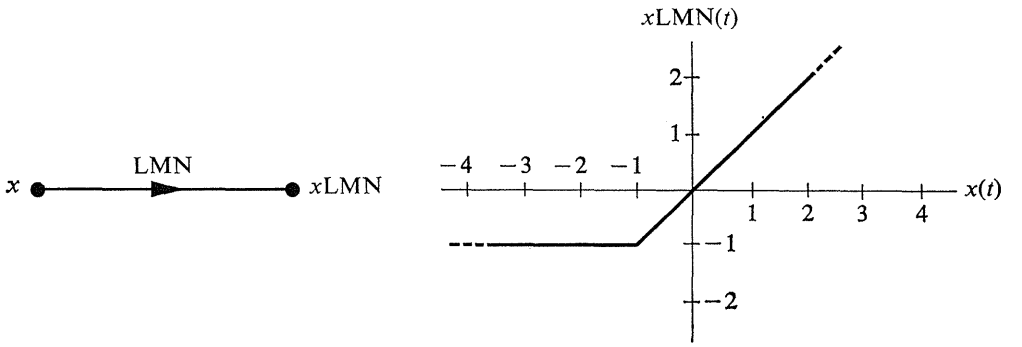


FIGURE 4.40

than some value $-z$,

$$x\text{LMN}z(t) = \begin{cases} x(t) & \text{if } x(t) > -z, \\ -z & \text{if } x(t) < -z. \end{cases}$$

A *unit, symmetric limitor*, SLM, is defined by the input–output characteristics as illustrated by Figure 4.41. More generally, for limiting the output at the $\pm z$ levels, we may define the operator SLMz. As the sketch of the input–output characteristic shows, SLM has *unity* scalar transmittance for $|x(t)| < 1$, and zero *incremental* transmittance for

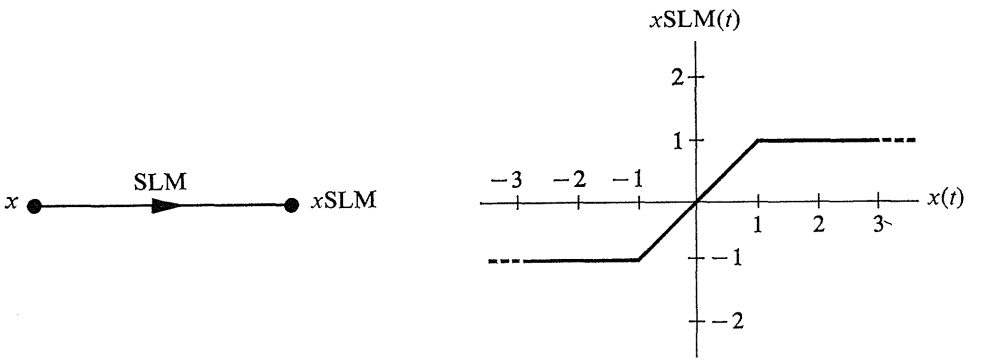
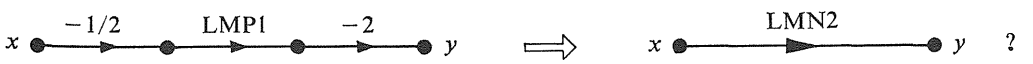


FIGURE 4.41

QUESTION 4.37 Is it true that



(Answer)

$|x(t)| > 1$. (A realization of this operator on the TR-10 analog computer is described on page 57 of the Operators Manual. It may be achieved (approximately) with the Personal Analog Computer by allowing the scalor box to overload.) A simple realization using a unit scalor and a DDZN operator is illustrated by Figure 4.42.

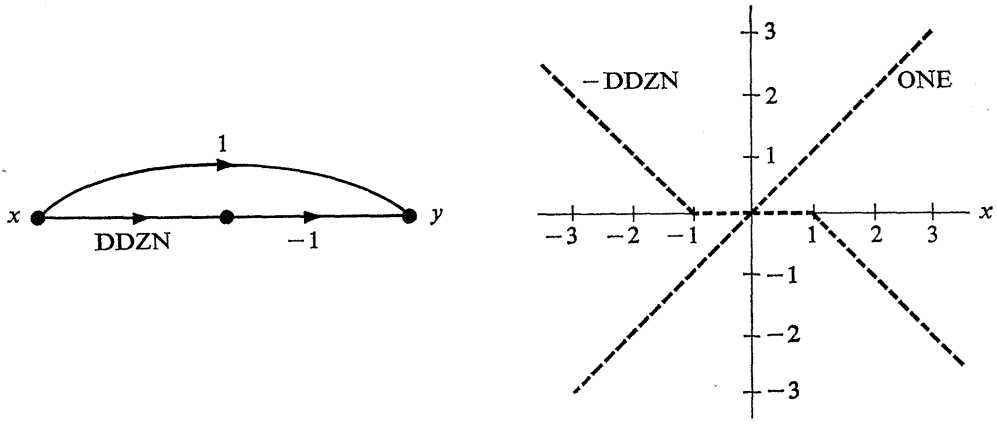
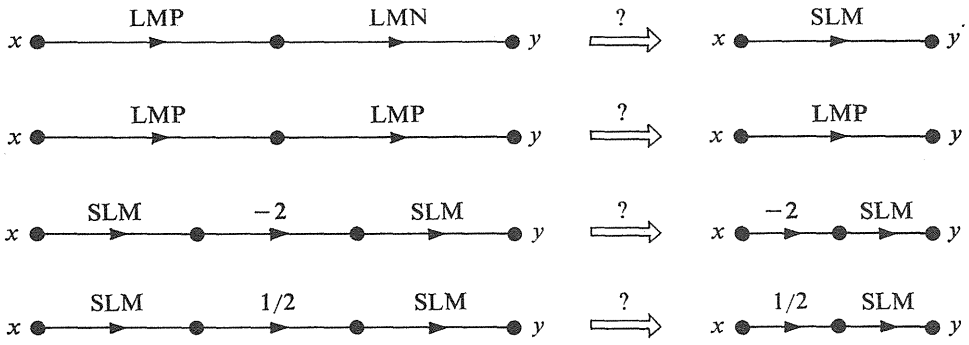
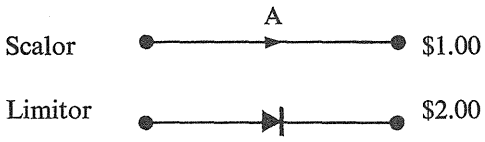


FIGURE 4.42

QUESTION 4.38 Which of the following implications hold true for all input signals x?

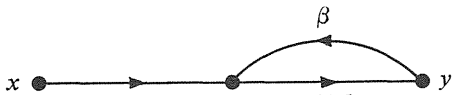


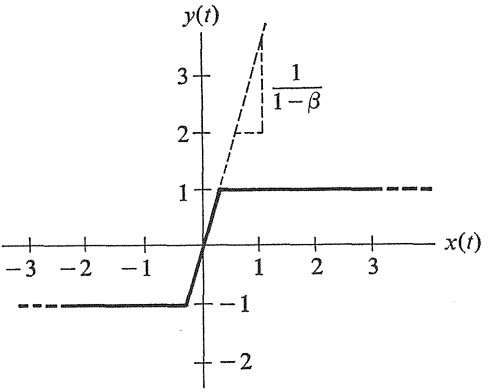
QUESTION 4.39 Given the following cost of parts:



What is the least expensive symmetrical limitor, SLM*z*, you can design?

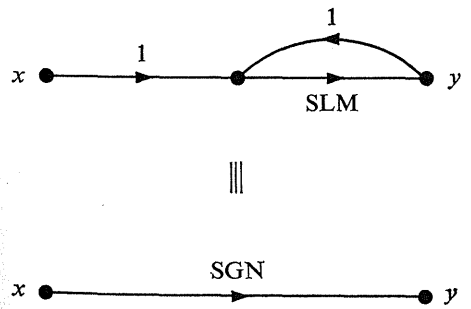
QUESTION 4.40 Show that the effect of *positive* feedback around a symmetrical limitor is to increase the slope of the *transition* region in its input–output characteristic.





(Answer)

Signum operator • An important special case of the operator considered in Question 4.40 occurs when $\beta = 1$. Then the width of the linear transition region is zero and the output has either the value $+1$ or -1, depending on whether the input value is positive or negative. This operator, therefore, preserves the algebraic sign of the input signal, but destroys all amplitude information. (In speech-communications equipment, the signum



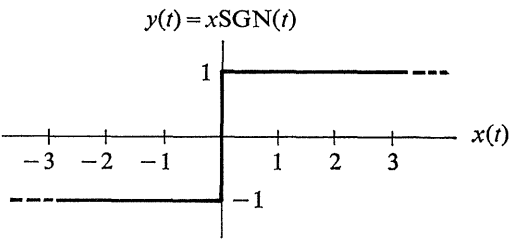


FIGURE 4.43 *Definition of the signum operator.*

operator, often called a *superlimiter*, is sometimes used to compress a speech signal so as to remove all variations in loudness. Although the processed speech signal is badly distorted, the *zero-crossings* are accurately preserved and, surprisingly, the distorted speech signal is still intelligible.) We shall denote this operator by SGN where

$$x\text{SGN}(t) = \begin{cases} 1 & \text{if } x(t) > 0 \\ 0 & \text{if } x(t) = 0 \\ -1 & \text{if } x(t) < 0 \end{cases}$$

The signum operator provides a convenient way for comparing the values of two signals. For instance, the output of the operator shown in Figure 4.44 will be $+1$ whenever $x(t) > y(t)$ and -1 whenever $x(t) < y(t)$. (We illustrate other uses of this

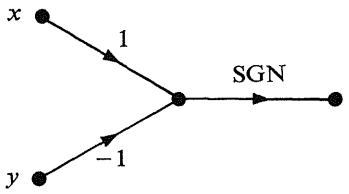


FIGURE 4.44

important operator in a later chapter.) Before concluding this discussion of various nonlinear operators realizable from limitors and scalors, we should mention one other operator of major importance that exhibits a novel property, not hitherto encountered. This is the property of *hysteresis.*

Hysteretic operator (or flip-flop) • Merely by adding unity positive feedback to the signum operator, we obtain a remarkable new property, illustrated by Figure 4.45. You

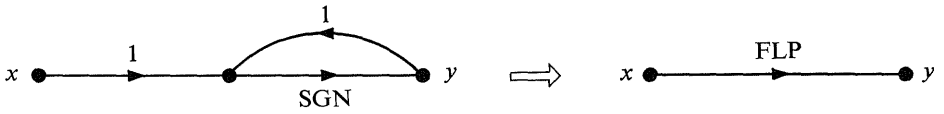


FIGURE 4.45

should verify that the input–output characteristic of this new operator is described by the plot of Figure 4.46. (Remember that the value of y is either $+1$ or -1.)

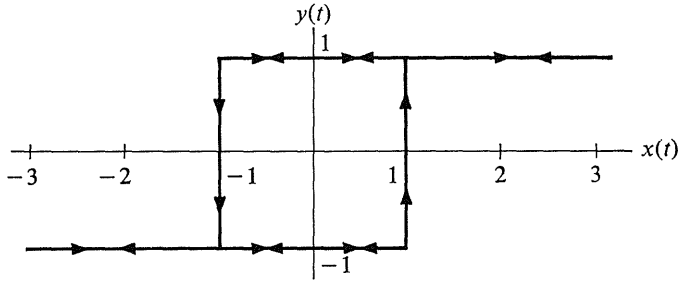
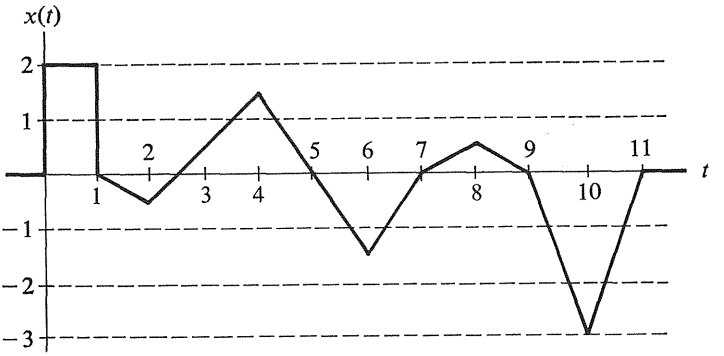


FIGURE 4.46 *Input-output characteristic of a flip-flop.*

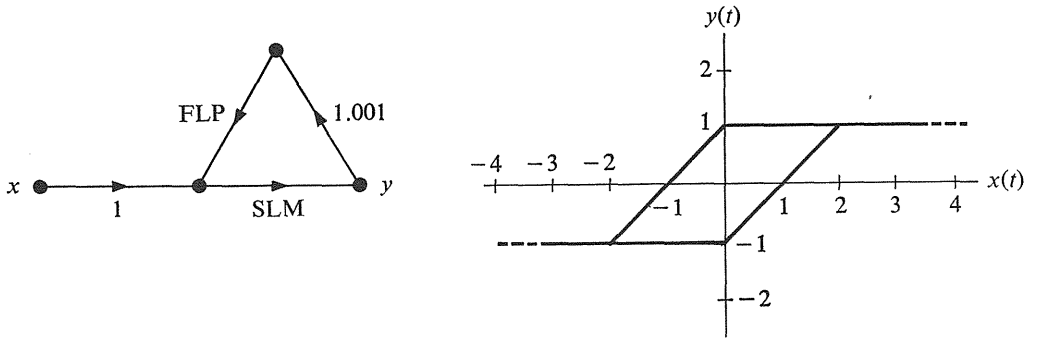
QUESTION 4.41 If, at some instant t, the value of the input is between -1 and $+1$, the input–output characteristic of the flip-flop shows that the value of the output may be either $+1$ or -1. What will determine which of these two possible output values actually occurs at that instant? (Answer)

QUESTION 4.42 Determine and plot, at each instant during the interval $0 < t < 10$, the output signal xFLP when the value of the input signal is as shown. (Answer)

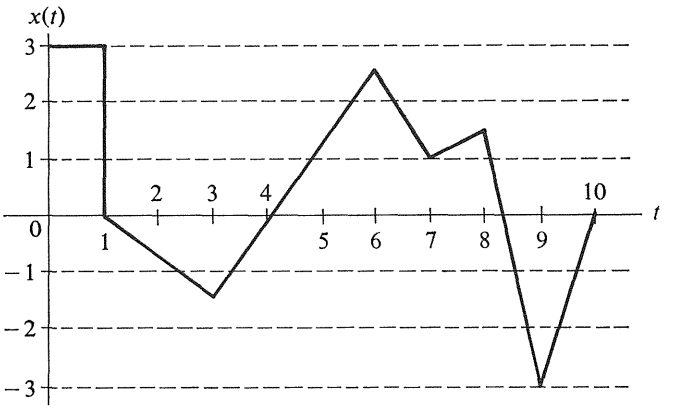


QUESTION 4.43

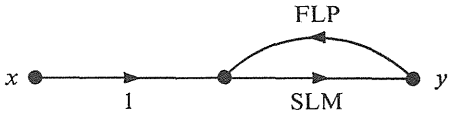
1. Verify that the input–output characteristic of the accompanying operator is as shown.



2. Plot $y(t)$ if $x(t)$ varies with time in accord with this graph.



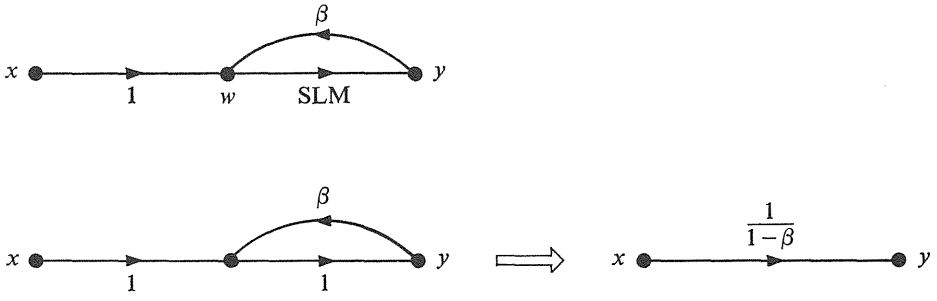
3. Would the accompanying operator graph have an input–output characteristic similar to FLP?



(Answer)

ANSWER TO QUESTION 4.37 Yes! This illustrates the fact mentioned earlier that the choice of a basic nonlinear operator is not unique. We may use one kind to construct many others, using only scalors as the second basic operator.

ANSWER TO QUESTION 4.40 When $|y(t)| < 1$, the value of the input signal w to SLM must be equal to $y(t)$. Thus, for $|y(t)| < 1$, SLM may be replaced by a scalor of unity transmittance, as shown. Hence, the portion of the input–output characteristic for $|y(t)| < 1$ is similar to that of a scalor of transmittance $1/(1 - \beta)$.



ANSWER TO QUESTION 4.41 The properties of the flip-flop illustrated by the problem suggest that it may provide a useful way of *storing information*. So long as the input signal has an absolute value that is less than unity, the flip-flop may be in either of two states: in the *up* state, with the value of its output signal equal to $+1$, or in the *down* state, with the value of its output equal to -1. To *switch* the flip-flop from the *up* state to the *down* state, the input signal must be *more negative* than -1. To *switch* the flip-flop from the *down* state to the *up* state, the input must be *more positive* than $+1$. The switching occurs instantaneously, and the flip-flop *remains indefinitely* in one or the other of its two states until the input signal achieves an appropriate value to cause it to change to the other state.

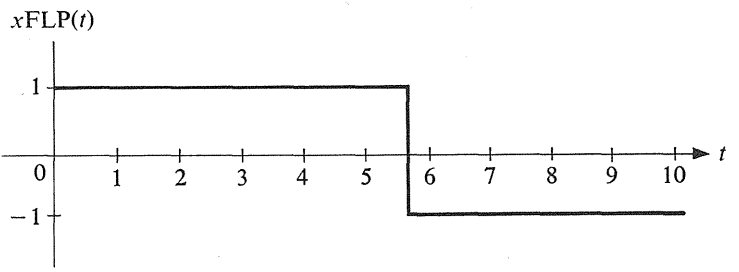
The input–output characteristic of a flip-flop exhibits a rectangular *hysteresis loop*. In fact, this nonlinear operator is an idealization of the essential characteristic of the magnetic-core memory element used in practically all modern computers, and in the recording of information on magnetic tape.

Another familiar illustration of this same flip-flop action is the ordinary toggle switch used to control the lights of a room. It also has ordinarily only two states, *on* and *off*, and the switch handle must be moved beyond a certain critical position to change the

switch mechanism from one state to the other. In fact, if one examines the mechanical linkage of a toggle switch, he finds: that it limits (because of mechanical *stops* that prevent the mechanism from moving beyond certain limits); that a spring provides a force which moves the mechanism away from its midposition; and that this force effectively acts as the *positive feedback* around SLM to drive the mechanism to its extreme position at either limit and hold it there until an "input" causes it to switch to the other state.

The flip-flop thus has a "memory" which does *not* depend directly on the *age* of the input. It provides a very useful element for *permanent* storage of information. The molecules of iron on a magnetic tape used in a sound recorder are, in effect, little flip-flops which can be placed in one state or the other by the recording process. By sensing the magnetic field associated with these flip-flops on playback, we are able to recover the information.

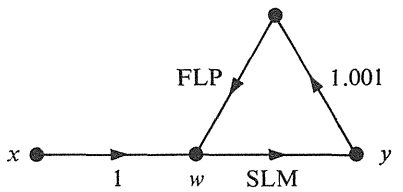
ANSWER TO QUESTION 4.42



ANSWER TO QUESTION 4.43

1. The state of the flip-flop FLP is dependent on the *most recent* value of x for which $|x(t)| > 1$. If the algebraic sign of this most recent value is positive, the flip-flop must presently be in the up state; if the algebraic sign is negative, FLP will be in the down state.

   Assume that FLP is in the up state. Then $w = x + 1$, and by SLM

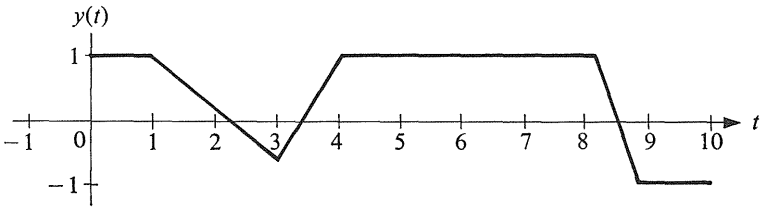


$$y = \begin{cases} x + 1 & \text{if } |x + 1| < 1 \\ 1 & \text{if } x + 1 > 1. \end{cases}$$

   This evidently accounts for the *upper* branch of the input–output characteristic.

   The input to FLP will be more negative than -1 when $y(t) < -0.999$, or when $x(t) < -1.999$. At the first instant when $x(t) < -1.999$, the flip-flop will flop to its down state and the value of w will suddenly become two units more negative (i.e., now $w = x - 1$ where as previously $w = x + 1$). This yields the *lower* branch of the input–output characteristic, where the operator remains until the value of the input exceeds $x(t) > 1.999$, when FLP flips to its up state, and the upper branch again applies.

2. The value of the output for the specified $x(t)$ is therefore



3. Without the scalor of transmittance slightly greater than 1, the input to FLP would never exceed 1 in magnitude and the state of FLP would not be alterable by $x(t)$.

From these several examples, it should be evident that nonlinear operators exhibit some remarkable properties. We have shown how to build up successively more complex operators, starting with only scalors and limitors, and how signal flow-graph notation may be used to describe these systems of operators. On the other hand, we have not been able to use Mason's rule to reduce the operators to simpler form. Such reductions are possible only when the individual operations are *linear*.

Comparison of Nonlinear with Linear Operations

If you will examine the five basic rules for interpreting signal flow-graphs given at the beginning of this chapter, you will see that they say nothing whatsoever about the linearity or stationarity of the individual operators. The signal flow-graph rules *do* require that the various signals flowing into each node *sum* to form a new signal at that node. This is a useful rule because useful signals *are additive*. Rule 3 is a simple convention for expressing a signal as the sum of other signals. Why then are we so interested in *linear* operators?

Linear operators are of great importance because the effect of two linear operators applied in succession to a signal is equivalent to a single linear operator acting on the signal. By inductive reasoning, starting from this property it is easy to show that any number of linear operations applied in succession is equivalent to a single linear operator. For instance, a signal in passing through a hi-fi amplifier is operated on by many different linear elements in the amplifier. The overall effect of the amplifier, which transforms the input signal into the output signal, is itself a linear operator. Because the individual elements are linear, they may be combined to find equivalent relations among the remaining signals.

Only with linear operators is reduction to an equivalent form possible, in general. For instance, the composite operator SQ^2, obtained by two successive applications of the squaring operator, is not a squaring operator. On the other hand, if L is a linear operator, then L^2 is itself linear. Because of this *semigroup** property of linear operators, whereby matings of linear operators beget only linear operators, the analysis of linear

* A set of operators is said to form a *group* if: the "product" of any two yields an operator that is already in the set; the set includes an identity operator; and the inverse of each operator exists and is also a member of the set. If only the first two conditions are satisfied (i.e., some of the operators may not have inverses belonging to the set), the set is called a *semigroup*.

systems is enormously simplified. Linear systems may be very large, but their analysis is not made more difficult because of this fact. The analysis only becomes longer.

The possibility of reducing a flow-graph is a direct consequence of its linearity. This is illustrated by operators shown in Figure 4.47. Here, A_1, A_2, B_1, and B_2 are scalors, but

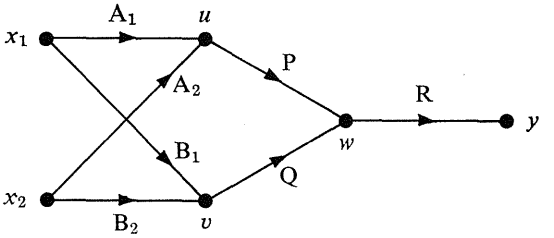


FIGURE 4.47 *General-operator graph.*

P, Q, and R are, for the moment, arbitrary operators. In algebraic symbols, this graph states that

$$(1) \quad x_1A_1 + x_2A_2 = u,$$

$$(2) \quad x_1B_1 + x_2B_2 = v,$$

$$(3) \quad uP + vQ = w,$$

$$(4) \quad wR = y.$$

If we wish to reduce this graph by eliminating node w, we could substitute (3) into (4) to obtain

$$(5) \quad [uP + vQ]R = y.$$

Unless R has the *additive* (i.e., distributive) property of a linear operator, further reduction of this expression is not possible. However, if R *is* linear, *then* this expression may be written as

$$(6) \quad uPR + vQR = y,$$

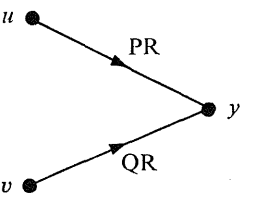


FIGURE 4.48

and node w in the graph of Figure 4.47 has been eliminated by expressing y directly in terms of equivalent operations on u and v. (Note that these equivalent operators need not themselves be linear.)

Next, suppose we wish to eliminate nodes u and v. Substitution of (1) and (2) into (6) yields

$$(7) \quad [x_1A_1 + x_2A_2]PR + [x_1B_1 + x_2B_2]QR = y,$$

and, again, further reduction is impossible unless P and Q are linear operators. Then,

$$x_1[A_1PR + B_1QR] + x_2[A_2PR + B_2QR] = y$$

or

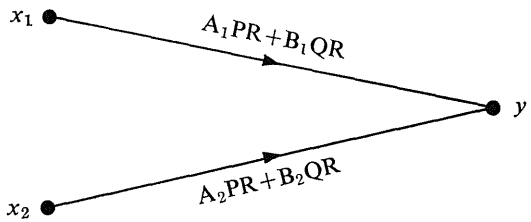


FIGURE 4.49

These are precisely the graph "transmittances" from each source node to the sink.

It should be evident that reduction of the graph to express the output directly in terms of the sum of equivalent operations on each of the input signals is possible *only* because the operators P, Q, and R are linear. This is why we are so interested in linear operators —only for systems of linear operators can we construct a reasonably simple *general* theory. (And, a substantial part of that theory is contained in the flow-graph methods that you have now learned.)

Commutative operators • Thus far we have not required that the operators *commute*. But, as pointed out in Chapter 2, unless they do commute, the algebraic reductions are more complicated. Furthermore, Mason's formula cannot be applied in its simple form to solve an operator graph.

Scalors are clearly commutative operators—the overall effect of AB is the same as BA on any signal, as illustrated by Figure 4.50. Scalors are *static* operators and not very

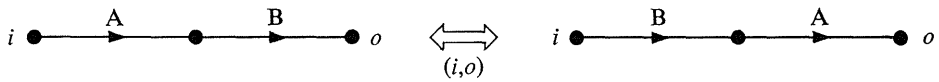


FIGURE 4.50 *Scalors are commutative operators.*

interesting. Physical systems are usually *dynamic* in the sense that the value of a signal at any instant depends on the values of other signals at earlier instants. The simplest *linear dynamic* operator that is *clearly* commutative is the *delayor*. It is, next to the scalor, the most fundamental operator.

A scalor yields a new signal whose value at each instant is a constant multiple of the value of the original signal at the *same* instant. A delayor yields a new signal whose value at time t is equal to the value of the original signal at the *earlier* time $t - T$, where T is the value of the time delay. Thus, suppose we applied in succession $\text{DEL}T_1$ and $\text{DEL}T_2$ to a signal x. The values of the signals at any instant t are illustrated by Figure 4.51. Evidently, two delayors in cascade are equivalent to a single delayor having a time

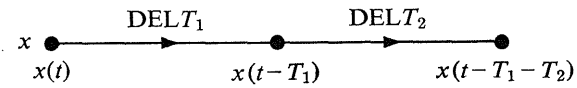


FIGURE 4.51 *Delays add.*

delay $T = T_1 + T_2$ equal to the *sum* of the two delays. From this it follows that *the same overall delay is achieved regardless of the order in which the delays occur.* So delayors commute with other delayors, and we have already shown that delayors and scalors commute. Delayors may therefore be included along with scalors in our system of operators, and we can apply flow-graph methods to them. It is remarkable that scalors and delayors provide primitive operators from which *all operations* found in *linear, stationary* physical systems may be constructed—no other kinds of operators are needed! All such operators, because they are composed of scalors and delayors, are likewise commutative. Our flow-graph methods therefore apply to the enormous class of linear, stationary physical systems that you will encounter in engineering work of all kinds.

Delayors

The overall effect of two cascaded delayors is equivalent to a single delayor whose time delay is the *sum* of the delay times of the constituent delayors. In contrast, two cascaded scalors are equivalent to a single scalor whose transmittance is the *product* of the transmittances of the constituent scalors. It might seem that delayors in cascade combine according to a "summation" rule, whereas scalors combine according to a "product" rule. By introducing a convenient notation, however, we may combine two cascaded delayors by the same "product" rule that applies to scalors.

Up to now, we have used the symbol DELT to denote a delayor whose delay is T. A slightly different notation will permit us to manipulate delayor symbols in exactly the same way as scalor symbols. The key idea is that *the value of the delay T appear as an exponent.* To support the exponent, we need some symbol to which it may be applied. We shall denote the delay operator by the composite symbol Z^{-T} where the "exponent," $-T$, signifies the time delay.* The fact that two cascaded delayors with delays T_1 and T_2, respectively, are equivalent to a single delay of time delay $T_1 + T_2$ is now automatically implied by our notation as illustrated by Figure 4.52. We may combine the cas-

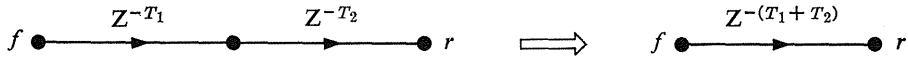


FIGURE 4.52 *Addition of delays is implied by the notation.*

caded delayors with a product rule because the same base, Z, is "raised" to the "powers" $-T_1$ and $-T_2$:

$$\mathrm{Z}^{-T_1}\mathrm{Z}^{-T_2} = \mathrm{Z}^{-(T_1+T_2)}.$$

As before, we combine this operator symbol with the symbol for the signal on which it acts to form the symbol for the resulting signal. If the value of the input signal f at the instant t is $f(t)$, then the value of the delayed signal $f\mathrm{Z}^{-T}$ at the instant t is $f\mathrm{Z}^{-T}(t) = f(t - T)$. Notice *no* delay at all (that is, $T = 0$) is represented by $\mathrm{Z}^{-0} = 1$, which is equivalent to a scalor of unity transmittance.

* In accord with present-day notation for Z-transforms (which are extensively used in the design and analysis of control systems), the choice of the symbol Z is quite arbitrary; we could have chosen DEL or any other symbol. However, the symbol Z does reveal the close ties with the theory of Z-transforms, and that is why we use Z.

An interesting question is whether the delay operator Z^{-T} has an inverse, $(Z^{-T})^{-1} = Z^{T}$. To be consistent, the symbol Z^{T} should denote an *anticipator* which would yield an output signal the value of which at any instant is the value of the input signal T time units *in the future*. Although such anticipatory operators are sometimes theoretically useful, *they are not physically realizable.* To correspond to a physical operator, the exponent of Z must not be positive. (You may ask, "Why then not define the delay operator with a positive exponent, instead of the negative exponent as in Z^{-T}?" The reason for the negative exponent is so that our results will agree with traditional Z-transform notation.)

Signal generators • With delayors in our stockpile of operators, we may now accomplish many additional operations. By using scalors and delayors, we may generate from a given signal new signals quite different from the original signal. (Scalors alone produce signals that are mere replicas of their inputs.) This possibility greatly extends the class of signals that may be described by our theory. By using interconnected systems of scalors and delayors, we may generate very complex signals from even the simplest signal.

To illustrate these possibilities, consider the very simple signal with unit value over the time interval $0 \leq t < 1$ and zero value elsewhere. The signal is like an idealized telegraph dot. We suppose that it is produced by a dot generator, shown in Figure 4.53.

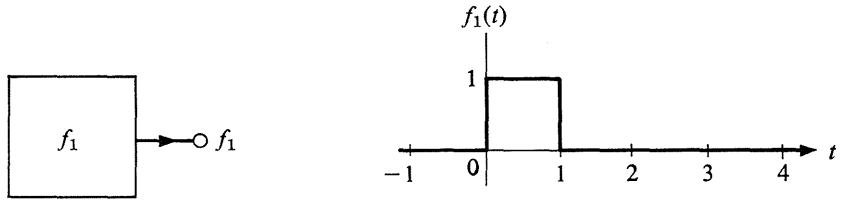


FIGURE 4.53 *A "dot" signal generator.*

How may we operate on this dot to produce a dash? The dash has unit value over the time interval $0 \leq t < 3$ and a value of zero elsewhere. A plot of the values of the dash f_2 as a function of time is shown in Figure 4.54. A dash is equivalent to three dots, of

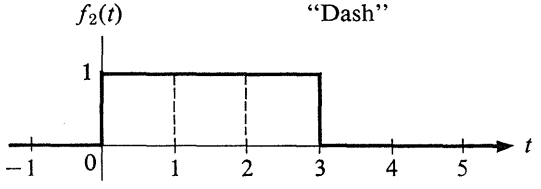


FIGURE 4.54 *A "dash" is composed of three "dots."*

which two are delayed. The second dot begins where the first ceases, and the third dot begins where the second ceases. Hence, we may generate a dash from a dot by applying the sequence of operations shown in the operator graph in Figure 4.55. We have thus

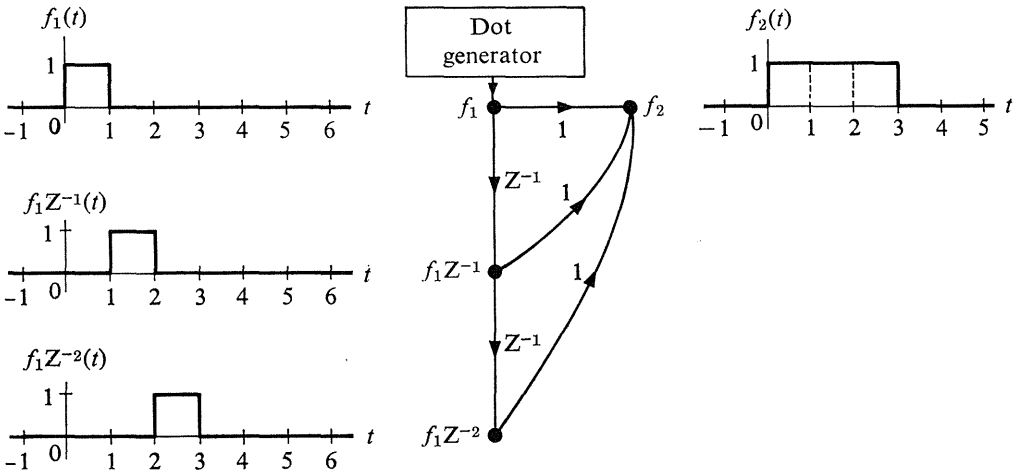


FIGURE 4.55 *Synthesis of a "dash" from a "dot."*

used the dot generator along with two delayors and scalors to make a dash generator, as shown in Figure 4.56.

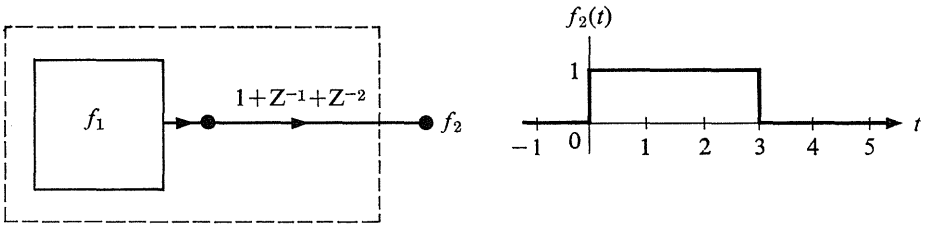
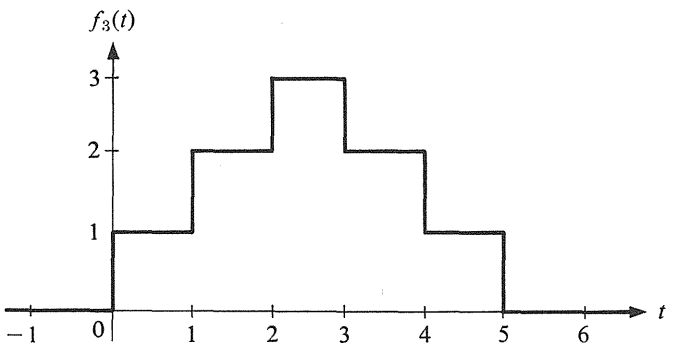


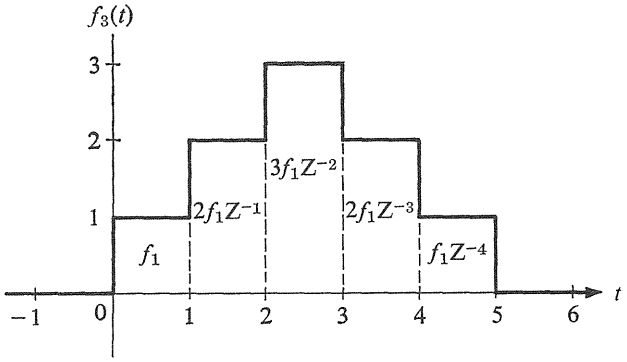
FIGURE 4.56

QUESTION 4.44 It is desired to generate the signal f_3, whose value at any time varies in accord with this plot. Specify, by drawing an operator graph, the operation upon the dot signal that will produce f_3. (Answer)



QUESTION 4.45 With reference to the same signal described in the preceding question, specify, by drawing an operator graph, the operation upon the dash signal f_2 that will produce f_3. Also, use the fact that $f_1(1 + Z^{-1} + Z^{-2}) = f_2$ to show that the operation just prescribed on f_2 corresponds to the *same* operation specified in Question 4.44 on the dot signal. (Answer)

ANSWER TO QUESTION 4.44 The desired new signal f_3 may be composed from delayed and amplified dots, as is evident from this sketch.



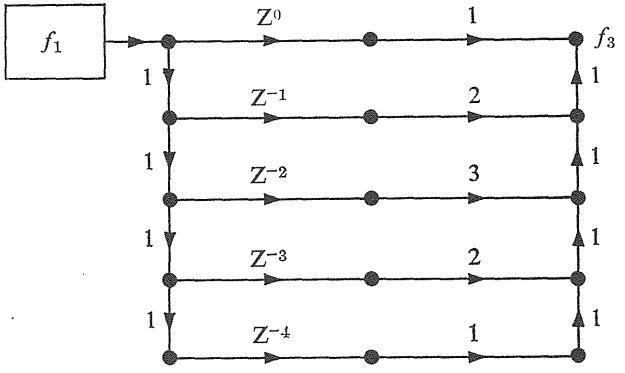
Hence,

$$f_1 + 2f_1Z^{-1} + 3f_1Z^{-2} + 2f_1Z^{-3} + f_1Z^{-4} = f_3$$

or

$$f_1[1 + 2Z^{-1} + 3Z^{-2} + 2Z^{-3} + Z^{-4}] = f_3.$$

In flow-graph form, we have a realization of the operator for transforming f_1 into f_3.



This operator graph shows that the value of f_3 at any instant t is given by the value of f_1 at the same instant multiplied by the scalar 1, plus the value of f_1 at the instant $t - 1$ multiplied by the scalar 2, plus the value of f_1 at the instant $t - 2$ multiplied by the scalar

3, etc. This operator is an example of the enormously important class of operators of the general form

$$\mathrm{H} = \sum_k a_k \mathrm{Z}^{-\tau_k}, \tag{4.11}$$

where an arbitrary scalar a_k is associated with each delay value τ_k.* We shall show later that by letting the number of terms increase without limit so that the difference, $\Delta\tau_k = \tau_{k+1} - \tau_k$, between successive delay values becomes vanishingly small, this general form encompasses the entire class of linear stationary operators relevant to physical systems.

When an operator H of the form given by Equation 4.11 is applied to any signal f, the resulting signal is composed of weighted time-shifted replicas of the original signal:

$$\begin{aligned} f\mathrm{H} &= f\left[\sum_k a_k \mathrm{Z}^{-\tau_k}\right] \\ &= \sum_k a_k f\mathrm{Z}^{-\tau_k}. \end{aligned} \tag{4.12a}$$

For the specific operator shown in the flow-graph, the values of τ_k are 0, 1, 2, 3, and 4, and the associated values of the weighting factors a_k are 1, 2, 3, 2, and 1. The value of the signal $f\mathrm{H}$ represented by Equation 4.12a at any instant t is

$$f\mathrm{H}(t) = \sum_k a_k f\mathrm{Z}^{-\tau_k}(t) \equiv \sum_k a_k f(t - \tau_k). \tag{4.12b}$$

Equation 4.12b clearly shows that the value of the output signal $f\mathrm{H}(t)$ depends on *earlier values* of the input signal, each *weighted* by an associated scalar factor. Thus, for the specific operator shown in the flow-graph,

$$\begin{aligned} f\mathrm{H} &= f[1 + 2\mathrm{Z}^{-1} + 3\mathrm{Z}^{-2} + 2\mathrm{Z}^{-3} + \mathrm{Z}^{-4}] \\ &= f + 2f\mathrm{Z}^{-1} + 3f\mathrm{Z}^{-2} + 2f\mathrm{Z}^{-3} + f\mathrm{Z}^{-4}, \end{aligned}$$

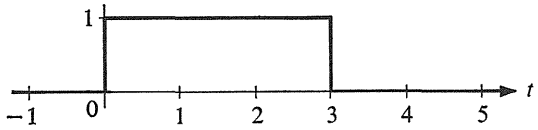
where we have written explicitly the individual terms in the summations indicated in Equation 4.12a. The value of the signal $f\mathrm{H}$ at any instant t is given by the sum of the values of its components at that instant:

$$f_3(t) = f_1(t) + 2f_1(t - 1) + 3f_1(t - 2) + 2f_1(t - 3) + f_1(t - 4).$$

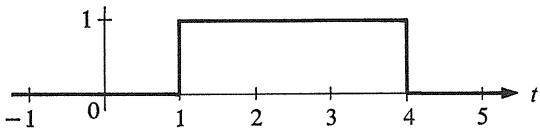
When $f_1(t)$ is the dot function portrayed in Figure 4.53, the function $f_3(t)$ takes on the specified values.

* The large Greek sigma is a standard symbol for denoting the sum of a set of terms, each having the form of the expression written after the sigma. This general form will involve an index (in this instance, k), and the variables (such as a and τ) are functions of this index. Traditionally, the index ranges over some set of integers, usually defined by writing the largest above the sigma sign and smallest below. When the index range is not of interest, we simply show the summation index.

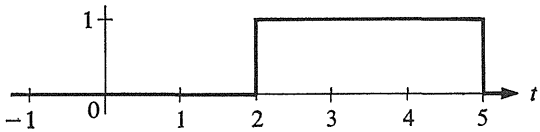
ANSWER TO QUESTION 4.45 The signal f_3 may also be composed from delayed dashes, as is evident from this sketch:



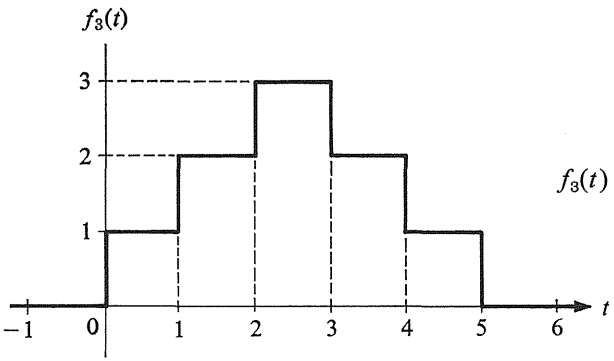
$f_2(t)$,



$f_2Z^{-1}(t)$,

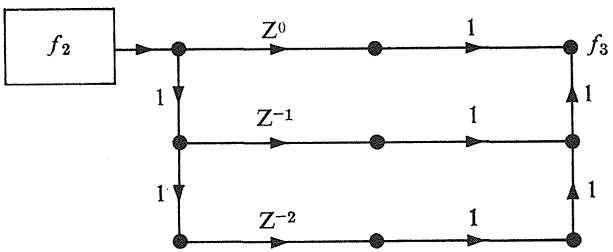


$f_2Z^{-2}(t)$.

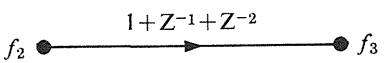


$$f_3(t) = f_2(t) + f_2Z^{-1}(t) + f_2Z^{-2}(t)$$
$$= f_2[1 + Z^{-1} + Z^{-2}](t).$$

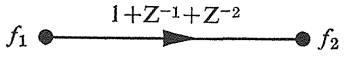
Expressing this composition by means of a flow-graph, we have



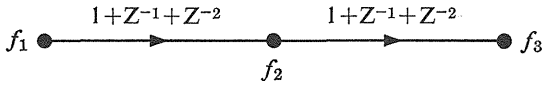
or



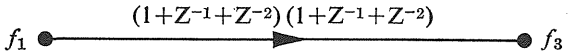
We have previously shown that the dash is itself obtainable by a suitable operation on a dot:



But, by the previous answer, f_3 may be obtained by an appropriate operation on f_2. Hence, to express f_3 in terms of f_1, we may combine the two linear operations



The overall result is found by ordinary "multiplication" of the two cascaded operators

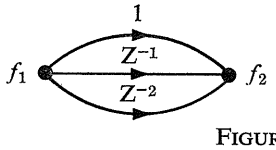


$$
\begin{array}{l}
1 + Z^{-1} + Z^{-2} \\
1 + Z^{-1} + Z^{-2} \\
\hline
1 + Z^{-1} + Z^{-2} \\
\quad + Z^{-1} + Z^{-2} + Z^{-3} \\
\qquad\quad + Z^{-2} + Z^{-3} + Z^{-4} \\
\hline
1 + 2Z^{-1} + 3Z^{-2} + 2Z^{-3} + Z^{-4}
\end{array}
$$

But this product is identical to the operator found in the previous problem. This example reveals how the operators may be manipulated algebraically and combined to get new results.

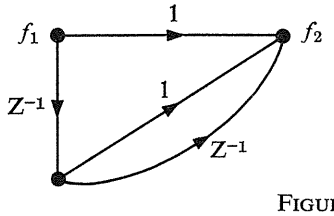
You can see why this particular notation is a convenient one for treating delayors.

The answers to Questions 4.44 and 4.45 show that often several different arrangements of scalors and delayors will produce the same result. The particular arrangement that is preferred will depend on engineering considerations such as, which one may be realized physically most cheaply, or most reliably, or most compactly. Physical realizations of delayors are usually quite expensive, with the cost being proportional to the delay time. Scalors are relatively cheap. To illustrate, suppose that each delayor costs \$10 per unit time delay or any fraction thereof (e.g., the delayor $Z^{-1.5}$ costs \$20), and suppose that scalors cost \$1 each. Now, to generate a dash from a dot, we may use the arrangement depicted in either Figure 4.57 or Figure 4.58.



Cost: \$ 1.00 scalor
30.00 delayors
\$31.00 total.

FIGURE 4.57



Cost: \$ 2.00 scalors
20.00 delayors
\$22.00 total.

FIGURE 4.58

If least cost were the criterion for choice, the second realization would be preferred. It is interesting to note that the first delayor does double duty in that realization, thus saving \$10 (of which \$1 is used for the second scalor). This possibility of arranging one operator strategically to accomplish the same result as many operators located elsewhere was mentioned in Chapter 2. There, a single phase equalizer at the transmitter of a television station was used rather than a separate phase equalizer in every television receiver.

The arrangement shown in the flow-graph in the answer to Question 4.44 is grossly inefficient and uses unnecessary components. If Z^{-0} is a delayor having a delay time of less than unity and therefore costing \$10, the realization of the operator graph shown in the graph would cost \$123. By removing unnecessary branches and rearranging the delayors in cascade to use the smaller delays in building up larger delays, we can save money.

QUESTION 4.46 John Jones claims that he can construct an operator for generating f_3 from a single dot f_1 for a total cost of parts of only \$44. How does he do it? (Answer)

Effect of loops • In our discussion of delayors, we have not thus far encountered feedback loops. Loops are widely used in control and communications systems. Phonograph records of rock-and-roll singers would sound worse without the delayors in feedback loops that enrich their renderings with artificial reverberation.

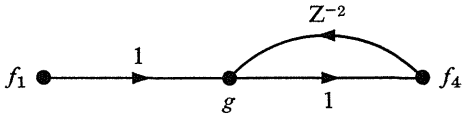


FIGURE 4.59

Let us investigate the signal f_4 that would be generated if the dot f_1 were applied to the operator graph with a loop as shown in Figure 4.59. This graph expresses f_4 in terms of the input signal f_1:

$$f_1 + f_4 Z^{-2} = g,$$

$$g = f_4.$$

Hence,

$$f_1 + f_4 Z^{-2} = f_4. \tag{4.20a}$$

At any instant t, the signals in Equation 4.20a have the values

$$f_1(t) + f_4(t - 2) = f_4(t). \tag{4.20b}$$

The value of the signal f_4 at any instant t is equal to the value of the input signal at that *same* instant plus the value of the signal f_4 at the instant $t - 2$, two time units *prior* to t.

Let us apply Equation 4.20b to find the value of $f_4(t)$ when $f_1(t)$ is as sketched in Figure 4.60. We assume that $f_1(t)$ is identically zero for all $t < 0$. Since f_1 is the *cause* of f_4, the value of $f_4(t)$ must likewise be zero for all $t < 0$. Consider then the value of $f_4(t)$ for

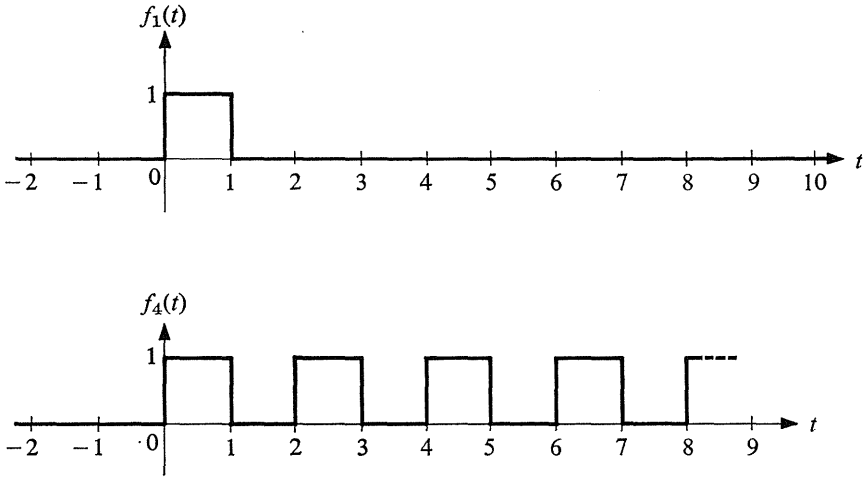


FIGURE 4.60

$0 \leq t < 2$. Over this interval, $t - 2$ will be less than 0. Hence, $f_4(t - 2)$ is zero and will not affect the value for $f_4(t)$, given by Equation 4.20b,

$$f_1(t) = f_4(t) \qquad \text{for} \qquad t < 2. \tag{4.21a}$$

Equation 4.21a enables us to plot the value of $f_4(t)$ for all $t < 2$. Over the time interval prior to $t = 2$, $f_4(t)$ looks just like $f_1(t)$.

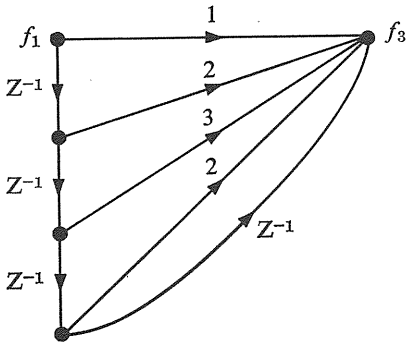
Next, consider the value of $f_4(t)$ during the time interval $2 \leq t < 4$. Over this interval, $f_1(t)$ is zero and Equation 4.20b simplifies to

$$f_4(t - 2) = f_4(t). \tag{4.21b}$$

But the value of $f_4(t - 2)$ is just the value of $f_4(t)$ two time units earlier, and this has already been found for any t less than 2. Therefore, the values of f_4 within the interval, $2 \leq t < 4$, are identical to the values at corresponding instants within the first interval $0 \leq t < 2$. Since the function $f_4(t)$ exhibits a dot at the beginning of the interval $0 \leq t < 2$, it will also exhibit a dot at the beginning of the interval $2 \leq t < 4$. By the same reasoning, the dot will be repeated at the beginning of the intervals $4 \leq t < 6$, $6 \leq t < 8$, etc. In other words, a delayor in a feedback loop has created a *periodic* train of pulses from a single pulse.

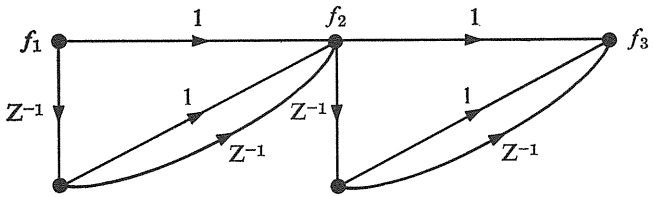
Delayors connected in feedback loops like that just illustrated are used extensively in modern communications equipment, radars, and digital computers. By repeating the input signal over and over at later times, they offer a kind of "memory." (Indeed, there is some evidence that closed loops of nerve cells may function in a similar way to provide short-term memory in our nervous system.) In the original UNIVAC digital computer,

ANSWER TO QUESTION 4.46 Two possible realizations for the operator needed to convert the dot f_1 into the stile-like signal f_3, are shown here.



Cost: \$ 4.00 scalors
40.00 delayors
\$44.00 total.

and



In the first graph, the equivalent algebraic form is

$$H = 1 + Z^{-1}\{2 + Z^{-1}[3 + Z^{-1}(2 + Z^{-1})]\}.$$

By factoring out the successive delayors, the total cost of the delayors is reduced to its minimum value. Note that this factoring technique is a systematic way of causing each factored operator to be utilized most fully by all subsequent operators. Thus, important design principles correspond to simple algebraic processes, such as factoring, that you already know.

You will find it useful to review some of these algebraic manipulations. In particular, given algebraic polynomials $P(x)$ and $D(x)$, with $P(x)$ of higher degree than $D(x)$, you may express $P(x)$ as

$$P(x) = D(x)Q_1(x) + R(x). \tag{4.13}$$

Equation 4.13 is more commonly written as

$$\frac{P(x)}{D(x)} = Q_1(x) + \frac{R(x)}{D(x)}. \tag{4.14}$$

Here, $Q_1(x)$ and $R(x)$ are the *quotient* and *remainder* polynomials obtained by "dividing" $P(x)$ by $D(x)$. The degrees of these polynomials satisfy the relations

$$\deg Q_1 = \deg P - \deg D, \tag{4.15a}$$

$$\deg R < \deg D. \tag{4.15b}$$

Given an algebraic polynomial, such as $P(x)$, we may assign a numerical value, for example a, to the symbol x and evaluate the numerical polynomial $P(a)$ for $x = a$.* If the value of $P(a)$ is zero, a is said to be a *zero* of the polynomial $P(x)$. The simplest polynomial which has a as its zero is $x - a$. This polynomial is of the first degree and it has only one zero (namely, a). More generally, a polynomial of the nth degree will have n zeros.

Let a be one of the zeros of $P(x)$. Let $D(x) = x - a$ be the simplest polynomial with a as its zero. Then, Equation 4.13 must hold for any numerical value assigned to the symbol x. In particular, it must hold when $x = a$:

$$P(a) = D(a)Q_1(a) + R(a). \tag{4.16}$$

Since both $P(a)$ and $D(a)$ are zero, the value of $R(x)$ must likewise be zero when $x = a$. But, by Equation 4.15b, $R(x)$ must be of zero degree (i.e., a constant). Since this constant is zero when $x = a$, it must *be zero for all* x. Thus, $(x - a)$ is an *exact divisor* of $P(x)$ because the remainder $R(x) = 0$:

$$P(x) = (x - a)Q_1(x). \tag{4.17}$$

If we divide $P(x)$ by $(x - a)$, the value of the remainder will be zero. This is the *remainder theorem.* It offers a simple way of checking whether a given value is indeed a zero of $P(x)$. (Use this method to show that -3 is a zero of $x^3 + 3x^2 + 2x + 6$.)

Now $Q_1(x)$ is a polynomial of degree one less than $P(x)$. Let a_2 be a zero of $Q_1(x)$. Then, by the same argument just given, $Q_1(x) = (x - a_2)Q_2(x)$, or

$$P(x) = (x - a_1)(x - a_2)Q_2(x), \tag{4.18}$$

where $Q_2(x)$ is a polynomial of degree two less than $P(x)$. Continuing in this way, we finally express the original nth degree polynomial as the product of n first-degree polynomials formed from the n zeros of $P(x)$:

$$P(x) = c(x - a_1)(x - a_2) \ldots (x - a_n). \tag{4.19}$$

Evidently, the way to factor an nth degree algebraic polynomial into its first-degree factors is to treat x as a *numerical* variable and solve the equation $P(x) = 0$ for its n roots, $a_1, a_2, \ldots, a_n$. Except for the constant multiplier c, these roots completely define the right-hand side of Equation 4.19.

A polynomial may be factored in different ways which are especially useful for particular purposes. For instance, to evaluate a polynomial such as $P(x) = x^4 + 2x^3 + 3x^2 + 2x^1 + 1$ for a given value of x, it is easier to first factor out x repeatedly as

$$P(x) = x(x(x(x + 2) + 3) + 2) + 1.$$

(Remember that in evaluating a cluster of nested parenthetical terms, you start with the innermost term.) This form requires 4 additions and 3 multiplications to evaluate $P(x)$ for x equal to some specified value, whereas to evaluate the more familiar form $x^4 + 2x^3 + 3x^2 + 2x + 1$ one could use as many as 4 additions and 13 multiplications. The

* Observe that the symbol x need *not* originally denote a numerical quantity—it may be an operator such as Z^{-1}. However, there is no reason why we cannot substitute for the symbol x a numerical quantity, if we choose to do so.

rearrangement of the terms in an algebraic expression is therefore important for computational purposes as well in the design of efficient systems.

In the second graph, the equivalent algebraic form is expressed even more compactly,

$$H = [1 + Z^{-1}(1 + Z^{-1})]^2.$$

By multiplying out either expression, we find that H is the desired operator:

$$H = 1 + 2Z^{-1} + 3Z^{-2} + 2Z^{-3} + Z^{-4}.$$

Of these two forms, that of the second graph offers the advantage of requiring the least variety of parameter values. It uses only scalors of unity transmittance and consists of two *identical* subunits connected in cascade. It therefore would be easier to assemble and test, because each subunit may be tested independently. It is preferred over the first realization for this reason.

mercury delay lines were connected in feedback loops to store the numbers (encoded into long sequences of dots) and to perform some of the arithmetical operations.*

The procedure that we have just used to find f_4 may be formalized by using Equation 4.20b to find the value of $f_4(t - 2)$ simply by replacing t by $t - 2$ (this is certainly legitimate since t represents *any* instant, and for any choice t, the instant $t - 2$ is equally acceptable). This substitution yields

$$f_1(t - 2) + f_4(t - 4) = f_4(t - 2). \tag{4.22a}$$

By substituting this into Equation 20b, we obtain

$$f_1(t) + f_1(t - 2) + f_4(t - 4) = f_4(t). \tag{4.22b}$$

But Equation 4.22b is also valid for any t. By replacing t by $t - 4$, we find that

$$f_1(t - 4) + f_1(t - 6) + f_4(t - 8) = f_4(t - 4). \tag{4.22c}$$

Substitution of this expression for $f_4(t - 4)$ into Equation 4.22b yields

$$f_1(t) + f_1(t - 2) + f_1(t - 4) + f_1(t - 6) + f_4(t - 8) = f_4(t). \tag{4.22d}$$

* A mercury delay line exploits the fact that acoustic waves travel through a medium like mercury with a more or less fixed velocity. The digital signal, consisting of a long sequence of dots, is applied to a quartz transducer to generate a corresponding acoustic signal at one end of a long mercury-filled tube. The acoustic signal arrives greatly attenuated at the other end of the tube, delayed by whatever time it takes for the sound to propagate through the mercury. It is then converted by another quartz transducer into a weak electrical signal that may be reamplified and reshaped to recover the original signal, now delayed. Because the acoustic signal is much weakened by energy losses as it propagates through the tube, the maximum length of the mercury column cannot exceed a few feet. This therefore limits the maximum time delay that can be achieved in this way. However, several delayors may be connected in cascade to obtain an equivalent overall delay equal to the sum of the delay times of the individual delayors.

The memory developed for the UNIVAC computer used clusters of six tubes in a single, temperature-controlled, mercury-filled tank with six input–output amplifiers mounted externally to achieve the necessary overall time delay. According to the paper by Auerbach *et al.* ("Mercury Delay Line Memory Using a Pulse Rate of Several Megacycles," *IRE Proc.*, *37*, no. 8 (August 1949), pp. 855–861); this recirculating memory could process 5,000,000 dots (i.e., "bits") per second. You will find this early descriptive paper quite interesting.

Continuing in this way, we obtain an ever-lengthening series of $M + 1$ terms for $f_4(t)$ of the form (with $T = 2$),

$$\left[\sum_{k=0}^{M-1} f_1(t - kT)\right] + f_4(t - MT) = f_4(t). \tag{4.22e}$$

Suppose that we continue this process indefinitely, thus allowing the integer M to increase indefinitely. Under what conditions on $f_1(t)$ are we assured that the series (4.22e) will yield a well-defined value for $f_4(t)$?

A sufficient (although not a necessary) condition that the infinite series obtained from Equation 4.22e converge to the value of $f_4(t)$ is that the value of the input signal f_1 should be zero for all *t prior* to some epoch. (The *epoch* of a signal is the time at which it begins, which, for convenience, we assume is at $t = 0$.) Then, by the assumption of causality, the value of f_4 must *also* be zero for all t prior to this *same* epoch. But note that the *last* term in the series given by Equation 4.22e is $f_4(t - MT)$. Whatever value of t we select at which to evaluate f_4, this last term will be the value of f_4 at an instant MT time units *prior* to t. That is, as M increases without limit, $(t - MT)$ decreases toward large *negative values*. But remember that both $f_1(t)$ and $f_4(t)$ are zero for all time prior to some epoch instant (for instance, at $t = 0$). The value of $f_4(t)$ is therefore well defined for any finite value of t by the seemingly infinite series

$$f_1(t) + f_1(t - T) + f_1(t - 2T) + \cdots = f_4(t) \tag{4.22f}$$

because all terms beyond the Mth (for which $MT > t$) will have a value of zero. This series is consequently *actually finite* for any $f_1(t)$ that vanishes prior to some epoch.

Equation (4.22f) is a statement about the signal values at certain instants. What is the corresponding statement about the signal itself ? Evidently, the various terms in Equation 4.22f are the values at time t of the signals in the expression

$$f_1 + f_1\mathrm{Z}^{-T} + f_1\mathrm{Z}^{-2T} + f_1\mathrm{Z}^{-3T} + \cdots = f_4. \tag{4.22g}$$

But this expression may be written as a single equivalent operation upon f_1:

$$f_1[1 + \mathrm{Z}^{-T} + \mathrm{Z}^{-2T} + \mathrm{Z}^{-3T} + \cdots] = f_4. \tag{4.22h}$$

It is easy to verify that the infinite series $[1 + \mathrm{Z}^{-T} + \mathrm{Z}^{-2T} + \mathrm{Z}^{-3T} + \cdots]$ is the inverse of the operator $[1 - \mathrm{Z}^{-T}]$.

QUESTION 4.47 Show by direct "multiplication" that the infinite operator series $[1 + \mathrm{Z}^{-T} + \mathrm{Z}^{-2T} + \mathrm{Z}^{-3T} + \cdots]$ is the inverse of the operator $[1 - \mathrm{Z}^{-T}]$.

Hence, Equation 4.22h becomes

$$f_1[1 - \mathrm{Z}^{-T}]^{-1} = f_4 = f_1\left[\frac{1}{1 - \mathrm{Z}^{-T}}\right]. \tag{4.22i}$$

But this is just the solution that we would have obtained from the original operator graph by Mason's method! In fact, we could have written down Equation 4.22g directly by

following the flow of signal f_1 through this graph *over all possible paths.* Thus, one path (over the two cascaded scalars) leads directly into the node for f_4. However, additional paths go around the loop once, twice, . . . , numbers of times. There are an infinite number of paths from f_1 to f_4 because of the loop. The signal f_4 is the sum of the incoming signals over all of these paths,

$$f_1 + f_1 Z^{-T} + f_1 Z^{-2T} + \cdots = f_4, \tag{4.23}$$

which is precisely Equation 4.22g.

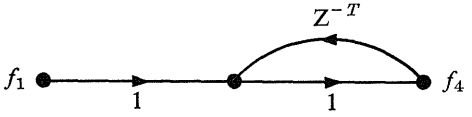


FIGURE 4.61

The principal result of the analysis involving Equations 4.20b through 4.22f is that *infinite expansions of delayors such as 4.23 will generally be valid provided that the value of the input signal f_1 vanish prior to some epoch.* This conclusion also applies to any operator graph composed of arbitrarily many delayors and scalors for precisely the same reasons discussed above.

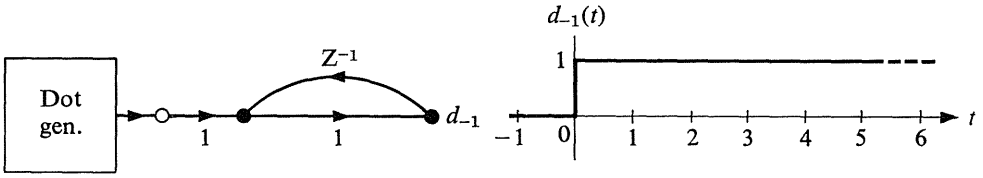


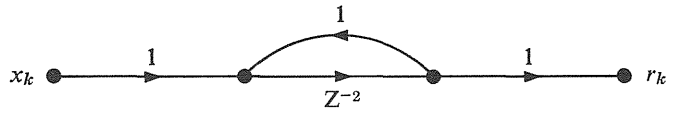
FIGURE 4.62 *Creation of a unit-step signal from a dot.*

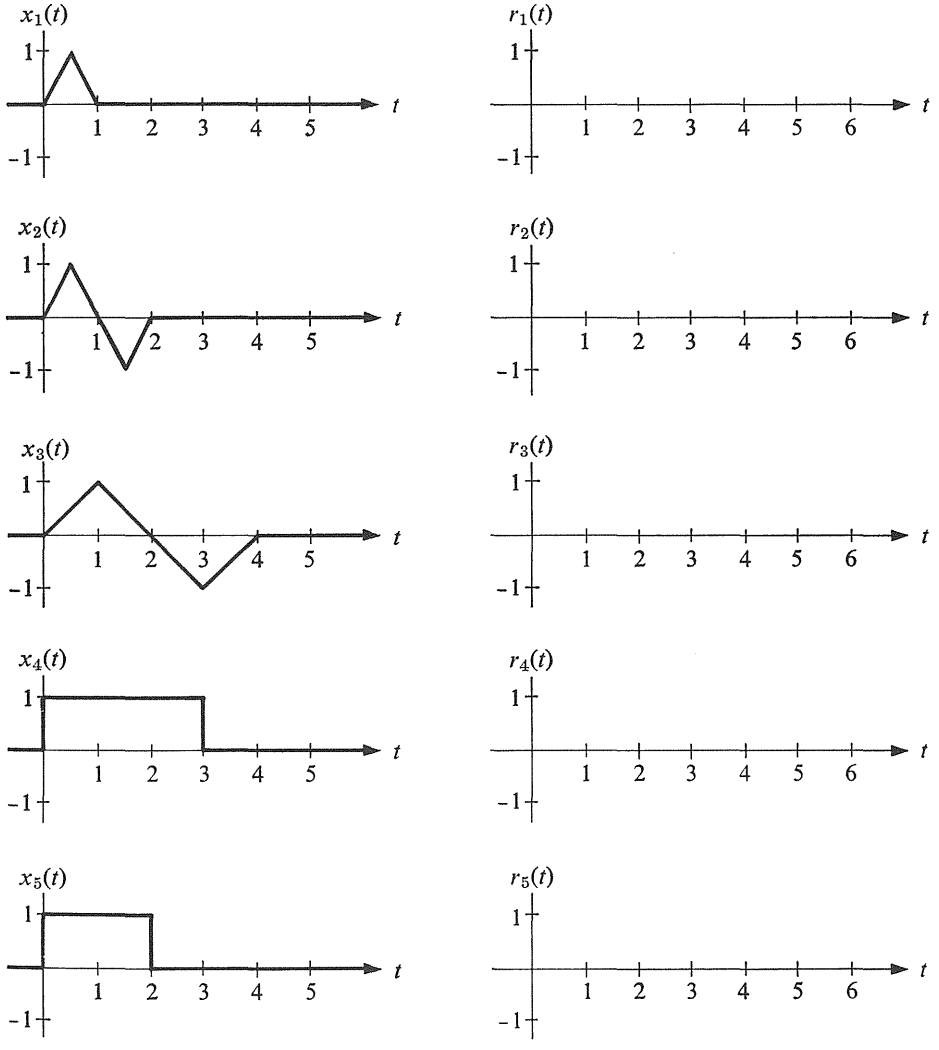
As a simple illustration of these ideas, we may generate a *unit-step* signal d_{-1} from a unit dot by means of the simple one-loop operator, as shown in Figure 4.62 where

$$d_{-1}(t) = \begin{cases} 1 & \text{if } t \geq 0, \\ 0 & \text{if } t < 0. \end{cases}$$

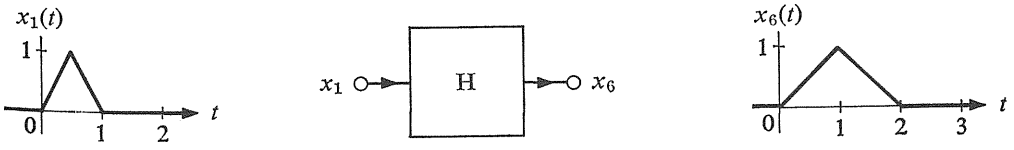
In a subsequent section, we shall see that the *unit-step function* $d_{-1}(t)$ plays an important role in describing signals created by physically realizable operators.

QUESTION 4.48 Plot the value at any instant within the time interval $-1 < t < 8$ of the output signal r_k when each of the different input signals x_k shown is applied to the operator graph. (Answer)



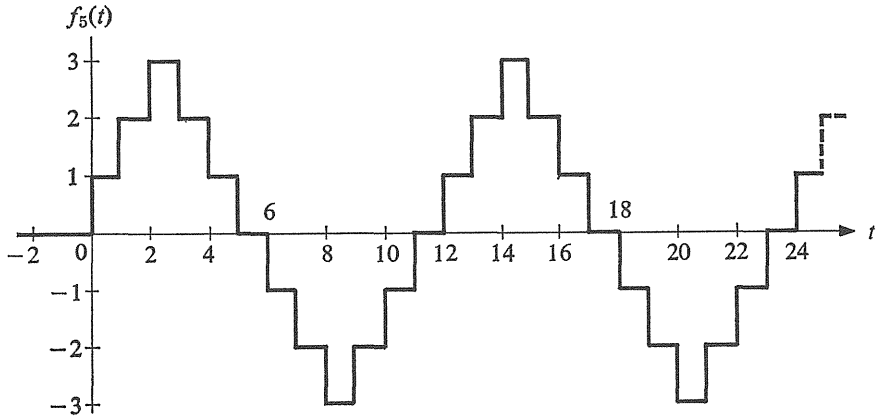


QUESTION 4.49 Given the triangular pulse x_1, design an operator H for transforming x_1 into a triangular pulse x_6 of twice the duration of x_1: (Answer)



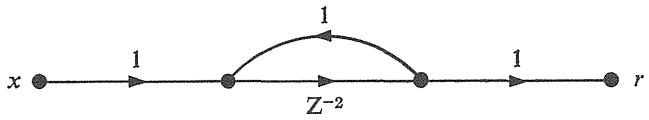
QUESTION 4.50 Suppose that the signal x_1 considered in the previous question were first recorded on magnetic tape moving at 7.5 in./sec, and immediately played back with the tape running at 3.75 in./sec. Would the played-back signal resemble x_6 (assuming a perfect recorder and playback device)? Would this process be linear? stationary? dynamic? (Answer)

QUESTION 4.51 Design a signal generator for the periodic signal f_5 in the accompanying sketch, using a dot f_1 as the original signal. What is the lowest cost that you achieve? (Answer)



QUESTION 4.52 In the $67 graph shown in the answer to Question 4.51, there are three scalor branches each with unity transmittance. Can any of these branches be eliminated without jeopardizing the performance of the operator? (Answer)

ANSWER TO QUESTION 4.48

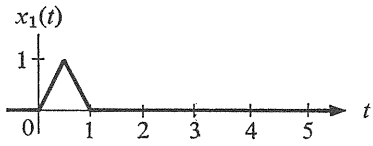


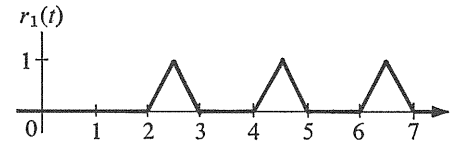
Here, the equivalent-operator series is

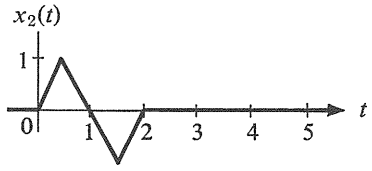
$$\frac{Z^{-2}}{1 - Z^{-2}} = Z^{-2} + Z^{-4} + Z^{-6} + Z^{-8} + \cdots .$$

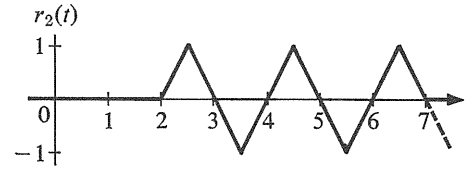
Hence, the r consists of the sum of time-delayed replicas of the input signal x,

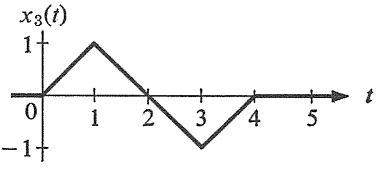
$$r = xZ^{-2} + xZ^{-4} + xZ^{-6} + xZ^{-8} + \cdots .$$

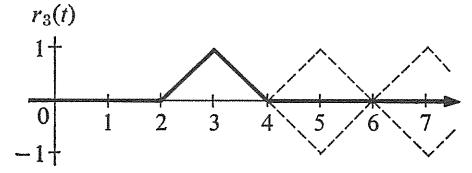


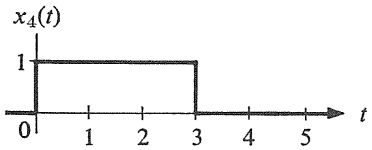


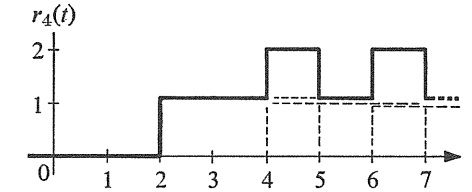


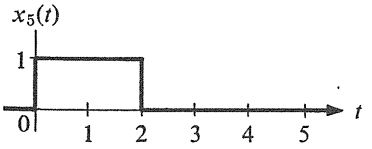


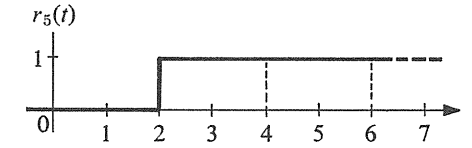




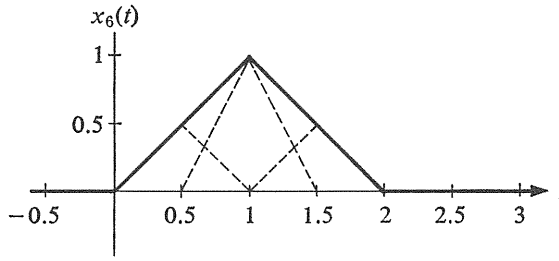








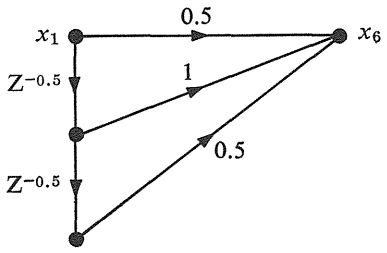
ANSWER TO QUESTION 4.49 We must first decompose x_6 into weighted and delayed replicas of x_1. The sketch shows how this may be done.



Hence,

$$0.5x_1 + x_1Z^{-0.5} + 0.5x_1Z^{-1} + x_6.$$

This operation may be accomplished by the following arrangement:

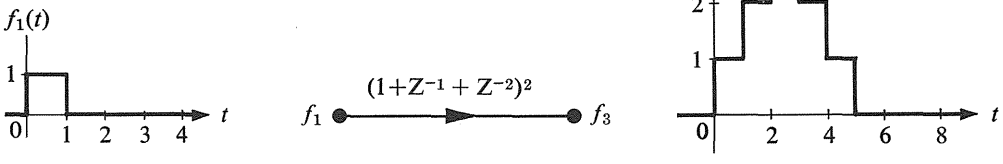


Cost: $23.00

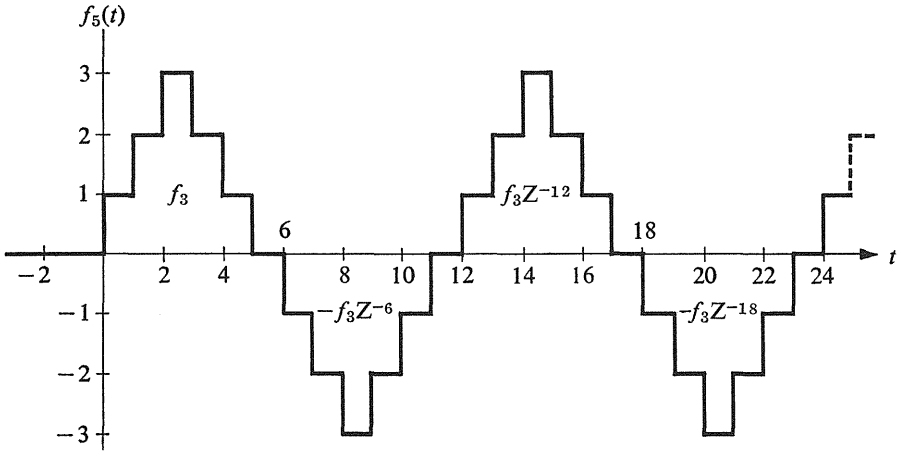
Note that a scalor changes only the *amplitude* of a signal at each instant and has *no effect* on the time axis. We might wonder whether a *time-stretch* operator is physically realizable. Question 4.50 suggests that it may be!

ANSWER TO QUESTION 4.50 In a magnetic tape recorder, the playback head is spaced about 1 in. farther along the tape than the record head, and it is standard practice to monitor what has just been recorded. What one hears must be delayed by at least 1/7.5 sec. However, by providing a take-up loop so as to increase the length of tape between the record and playback head, this time delay can be increased greatly. In fact, the length of the loop can vary with time, and interesting effects may be achieved in this way. In principle, this process is *linear*, although certainly not stationary. A very interesting operation is achieved by replaying the tape, but in the reverse direction. This *time-reversal* operation is ideally linear, but nonstationary. Signals which exist over finite time intervals may be reversed in time so that on playback what was originally the *past* now becomes the *future* and vice versa. This is a useful method for realizing "non-realizable" anticipatory operators. (Of course, the large time lag between the original occurrence of the signal and the "time-reversed" reproduction makes this operation consistent with the principle of physical causality.)

ANSWER TO QUESTION 4.51 In Question 4.44, we showed that it was possible to generate a stile-like signal f_3 from a dot f_1, as shown. Evidently the periodic



signal f_5 is composed of f_3 components of alternating sign, each delayed by an additional 6 time units, as shown below.



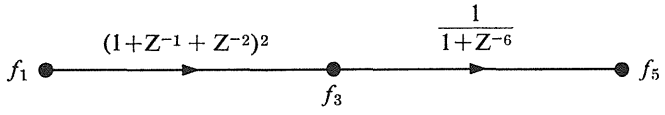
Hence,

$$f_3[1 - Z^{-6} + Z^{-12} - Z^{-18} + \cdots] = f_5$$

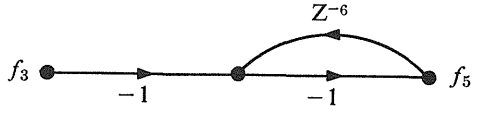
or

$$f_3\left[\frac{1}{1 + Z^{-6}}\right] = f_5.$$

This sequence of operations may be performed by the cascaded operators

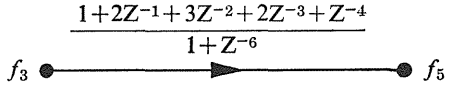


The first operator for generating f_3 from f_1 has already been designed by John Jones for a parts cost of \$44 (see Question 4.46). The second operator may evidently be realized by the graph

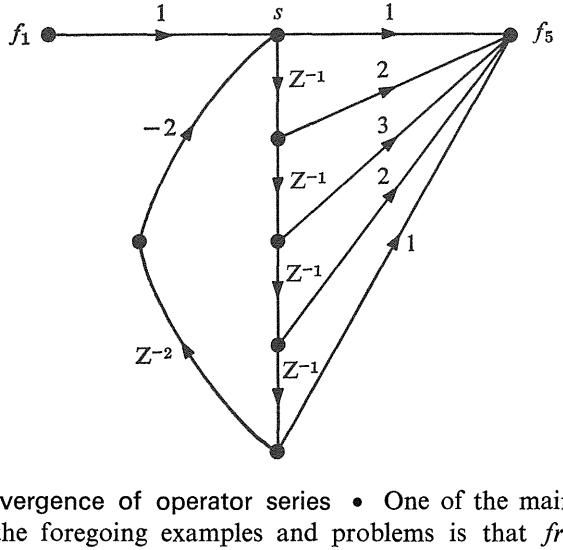


at a cost of \$62. Thus, it appears that the total cost should be \$44 + \$62 = \$106.

It is possible to reduce the complexity of the operator (and hence its cost) by combining the two parts just described. The overall operator for converting f_1 into f_5 is



This may be realized, with reference to Mason's rule, by the accompanying graph. By combining the two operators, we have been able to utilize the delayors more efficiently and reduce the total cost by 30%.



Cost: \$ 7.00 scalors
60.00 delayors
\$67.00 total

Convergence of operator series • One of the main points that has been illustrated by the foregoing examples and problems is that *fractional operator forms*, such as $1/(1 - aZ^{-1})$, *are merely an abbreviated notation for infinite series expansions in negative powers of* Z, such as

$$1/(1 - aZ^{-1}) \equiv 1 + aZ^{-1} + a^2Z^{-2} + a^3Z^{-3} + \cdots. \tag{4.24}$$

We have already shown several times that this power-series expansion is obtained by summing *all* possible paths through the associated graph. This provides an explicit interpretation for the operator $[1 - aZ^{-1}]^{-1}$ in terms of a sequence of well-defined operations.

It is worth noting that there is no restriction on the value of the scalar a except that it be finite. The convergence of the infinite series (4.24) depends upon the signal f to which it is applied as well as upon the operator.

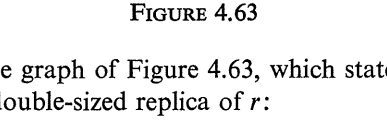


FIGURE 4.63

Consider for instance the graph of Figure 4.63, which states that the signal r is composed of f and a delayed double-sized replica of r:

$$f + 2rZ^{-1} = r.$$

To express r in terms of f, group terms in r,

$$f = r[1 - 2Z^{-1}],$$

and then operate on both sides by $[1 - 2Z^{-1}]^{-1}$,

$$f\left[\frac{1}{1 - 2Z^{-1}}\right] = r. \tag{4.25}$$

This is, of course, identical to the expression that we would have obtained directly by Mason's rule. Expansion into series form just described gives

$$f + 2fZ^{-1} + 4fZ^{-2} + 8fZ^{-3} + \cdots = r. \tag{4.26}$$

At any instant t, the values of the signals in Equation 4.26 are

$$f(t) + 2f(t-1) + 4f(t-2) + 8f(t-3) + \cdots = r(t). \tag{4.27}$$

Suppose that the input signal is the dot signal, for which $f(t) = 1$ for $0 \leq t < 1$, and is zero elsewhere. Then the plot of $r(t)$ will appear as in Figure 4.64. The value of the signal r will grow rapidly as time increases—in fact, with the passage of each unit of time,

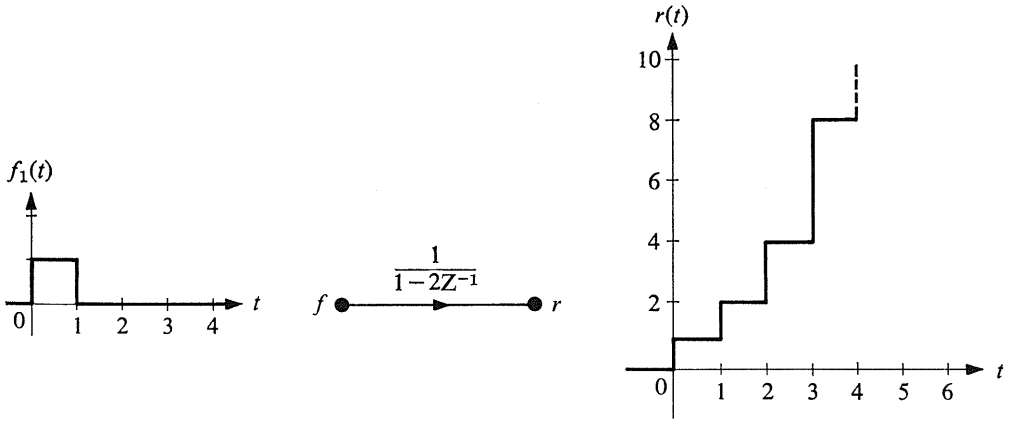


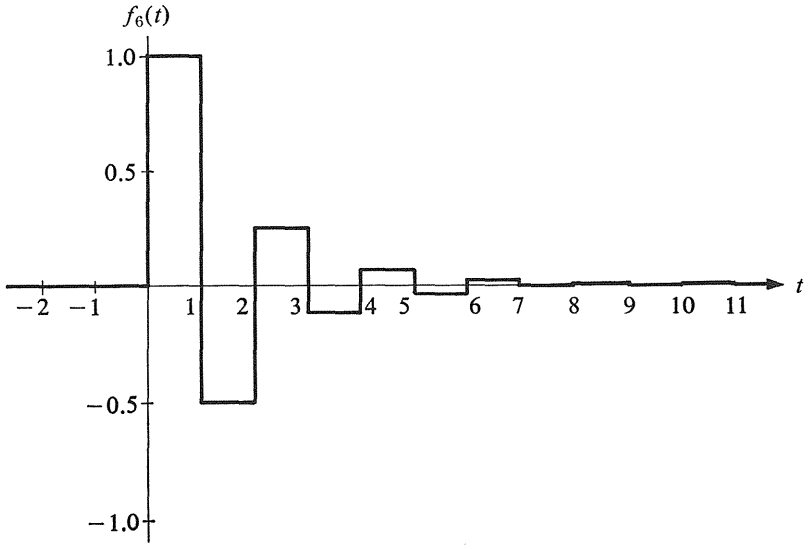
FIGURE 4.64 *A growing-signal generator.*

the value of r exactly doubles. However, the infinite series expressed by Equation 4.27 is perfectly well defined because it is actually a finite series by virtue of $f(t)$ being zero for all $t < 0$. In fact, for this particular $f(t)$, all terms except one in the series are zero for any $0 < t < \infty$, and that one term clearly has a well-defined finite value.

QUESTION 4.53 If, instead of a dot, suppose that the *unit-step* signal f_4 were applied to the operator just considered. A different response signal r will of course be obtained. What is the value $r(2.5)$ of the signal r at the instant $t = 2.5$? (Answer)

QUESTION 4.54 With regard to the same operator shown in Figure 4.64, suppose that in addition to the input signal shown there, a very small dot, having an amplitude of only 1/1,048,576, had occurred prior to $t = 0$ during the time interval $-20 \leq t < -19$. What effect would this tiny input signal have on the value of $r(t)$ during the interval $0 \leq t < 1$, as shown in Figure 4.64? (Answer)

QUESTION 4.55 Design an operator for generating the following signal, starting with a dot signal as input. (Perhaps this is the kind of operator that reverberates the rock and roller's recorded renderings.) (Answer)



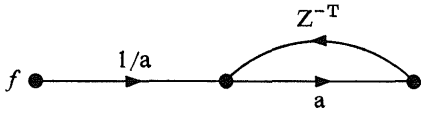
Stability • Having one or more feedback loops involving delay in an operator graph raises the question of *stability*. As time goes on, a signal may be repeatedly transformed by the loop operator an infinite number of times. Preceding questions pointed out that these endlessly repeated operations may produce an output signal whose value grows indefinitely with the passage of time. Even the smallest input signal applied at some remote time will produce an enormously large output signal if given sufficient time to "grow." An operator which produces an output signal increasing without bound after *some* input signal has been permanently reduced to zero is said to be *unstable*. Conversely, if for *any* input signal having bounded values, the output signal values are bounded, the operator is said to be *stable*.

System stability is a topic of great practical significance. By definition, the properties of a *linear* operator do not depend on how big the signals are—hence, in principle, the signals may *grow indefinitely*. Also, by definition, the properties of a stationary operator do not change with the passage of time. Hence, if a stationary operator is unstable *now*, it must always have been unstable and will continue to be unstable forevermore, unless the system is *nonlinear*. In this event the large signals may modify system behavior to make it stable when the signals exceed some limiting magnitude. Physical linear stationary operators are therefore always stable, for if they were not stable, they would have

ANSWER TO QUESTION 4.52 All three of the scalar branches are needed. Node f_1 is a *source* node and its value is specified independently of other signal values, whereas node s is equal to f_1 plus *additional signal* returned around the delayor loop. If node s were made to coincide with node f_5, spurious loops would be created.

to be either nonlinear or nonstationary. (It would be more accurate to state that a physical system is never completely linear nor stationary, but that over some prescribed interval of time and range of signal magnitudes, the system may be approximated by an appropriate linear stationary operator.)

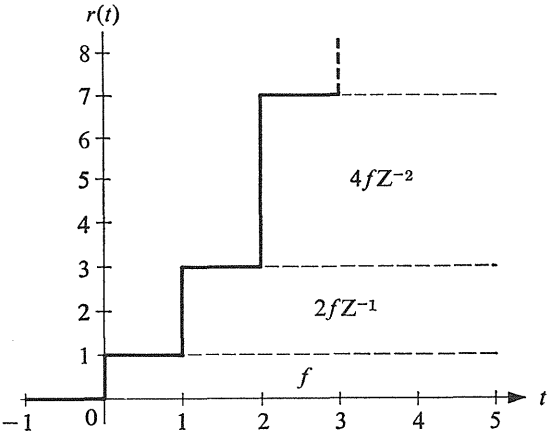
Although we shall investigate the stability question much more thoroughly in a later chapter, the results of Questions 4.53 through 4.55 already provide an answer to this question for any operator H comprised of a single delayor loop such as that shown in Figure 4.65. Here



$$H = \frac{1}{1 - aZ^{-T}}.$$

FIGURE 4.65

ANSWER TO QUESTION 4.53 The response now is composed of a succession of step-signal components, shown in the accompanying sketch. From this plot



it is clear that the value of $r(2.5)$ is 7. For this value of t, all terms in the series of Equation 4.27 beyond the first three terms are zero because $f(t) = 0$ for $t < 0$.

ANSWER TO QUESTION 4.54 We wish to find the response of the operator $[1 - 2Z^{-1}]^{-1}$ to a very weak "noise" dot n, where

$$n(t) = \begin{cases} \dfrac{1}{1{,}048{,}576} & \text{for } -20 \leq t < -19, \\ 0 & \text{elsewhere.} \end{cases}$$

The response of the operator to this very small noise dot is found, just as before, by evaluating

$$n + 2nZ^{-1} + (2)^2nZ^{-2} + \cdots + (2)^knZ^{-k} + \cdots.$$

The only signal component which is not zero during the interval $0 \leq t < 1$ is the term for which $k = 20$:

$$(2)^{20} nZ^{-20}(t) = (2)^{20} n(t - 20).$$

During the interval $0 \leq t < 1$, the value $n(t - 20)$ of the noise pulse is known to be 1/1,048,576. Hence, the value of the response during this interval is

$$(2)^{20} \frac{1}{1{,}048{,}576} = 1.$$

In other words, an extremely weak dot signal occurring 20 time units prior to the dot at $t = 0$ can produce an output signal which is identical subsequent to $t = 0$ to that produced by the unit dot at $t = 0$.

Can you suggest why this operator is said to be unstable? What practical difficulties do you foresee in using such an operator?

Each time it flows around the delayor loop, the signal is amplified by the scalar a, the nth circulation yielding the signal $a^n f Z^{-nT}$. Evidently, as n increases without limit, this term will vanish provided $|a| < 1$, whereas it will grow indefinitely if $|a| > 1$. When $|a| = 1$, we obtain a periodic signal which neither grows nor decays. Thus, *the operator will be stable provided* $|a| < 1$.

What may be said about the stability of complicated operators having two or more delayor loops? For instance, is the two-loop operator shown in Figure 4.66 stable? If a dot f_1 were applied at $t = 0$, what would be the value of the response signal r at any subsequent time?

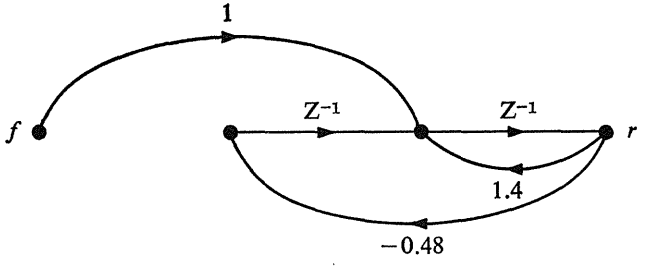
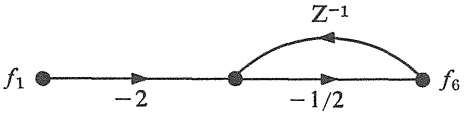


FIGURE 4.66 *A two-loop dynamic operator.*

ANSWER TO QUESTION 4.55 This question illustrates a basic property of a *stable* operator—that if after the input signal has been reduced to zero value, the output signal also dies away to zero value, the operator is surely stable. It is evident from the figure that

$$f_1 + (-1/2) f Z^{-1} + (+1/4) f Z^{-2} + (-1/8) f Z^{-3} + \cdots = f_6$$



Cost: $12.00

At first glance, it might appear that the operator shown in Figure 4.66 should be unstable because of the 1.4 transmittance feedback branch around the second delayor. However, it is not at all clear what canceling effect the outer feedback branch, of -0.48 transmittance, will have. A possible approach for determining the stability of this operator is simply to apply a dot signal and see how the response develops.

To do this, we find first the graph operator expressing r in terms of f:

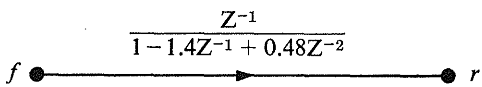


FIGURE 4.67 *Algebraic representation of the two-loop operator.*

Next, we expand this operator into a series involving negative powers of Z of the form $H = \sum_k a_k Z^{-k}$. This is easily done by ordinary long division:

$$
\begin{array}{r l}
 & Z^{-1} + 1.4Z^{-2} + 1.48Z^{-3} + 1.40Z^{-4} + 1.25Z^{-5} + 1.08Z^{-6} + 0.91Z^{-7} + \cdots \\
1 - 1.4Z^{-1} + 0.48Z^{-2} \;) & Z^{-1} \\
 & \underline{Z^{-1} - 1.4Z^{-2} + 0.48Z^{-3}} \\
 & \quad 1.4Z^{-2} - 0.48Z^{-3} \\
 & \quad \underline{1.4Z^{-2} - 1.96Z^{-3} + 0.67Z^{-4}} \\
 & \qquad 1.48Z^{-3} - 0.67Z^{-4} \\
 & \qquad \underline{1.48Z^{-3} - 2.07Z^{-4} + 0.71Z^{-5}} \\
 & \qquad\quad 1.40Z^{-4} - 0.71Z^{-5} \\
 & \qquad\quad \underline{1.40Z^{-4} - 1.96Z^{-5} + 0.67Z^{-6}} \\
 & \qquad\qquad 1.25Z^{-5} - 0.67Z^{-6} \\
 & \qquad\qquad \underline{1.25Z^{-5} - 1.75Z^{-6} + 0.60Z^{-7}} \\
 & \qquad\qquad\quad 1.08Z^{-6} - 0.60Z^{-7} \\
 & \qquad\qquad\quad \underline{1.08Z^{-6} - 1.51Z^{-7} + 0.53Z^{-8}} \\
 & \qquad\qquad\qquad 0.91Z^{-7} - 0.53Z^{-8}
\end{array}
$$

Thus, the response signal r of this operator is composed of scaled, delayed replicas of the input signal f as given by

$$
\begin{aligned}
r = f[Z^{-1} &+ 1.40Z^{-2} + 1.48Z^{-3} \\
&+ 1.40Z^{-4} + 1.25Z^{-5} \\
&+ 1.08Z^{-6} + 0.91Z^{-7} \\
&+ 0.76Z^{-8} + 0.62Z^{-9} \\
&+ 0.51Z^{-10} + 0.41Z^{-11} + \cdots]. \qquad (4.28)
\end{aligned}
$$

When the input signal f is the dot f_1, the value of the response signal is readily found to be as plotted in Figure 4.68. Inspection of the long-division process reveals that after about the tenth division the two remainder terms at each step are approximately $0.8Z^{-1}$ times the two remainder terms of the previous step. Thus, beyond about $t = 10$, the value of

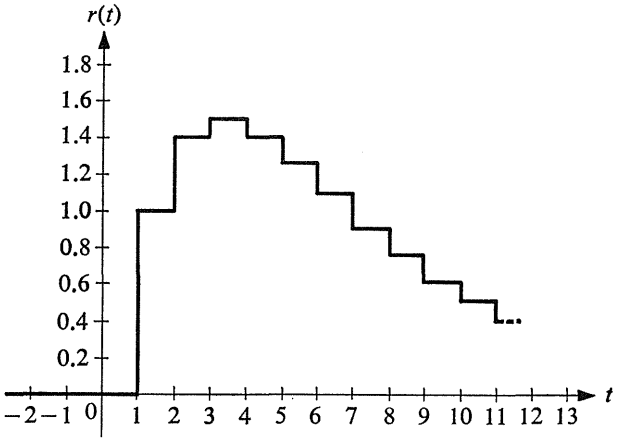


FIGURE 4.68 *The response of the two-loop operator of Figure 4.66 to a single dot input signal.*

the signal changes by a multiplicative factor of 0.8 for each additional unit of time. This operator therefore appears to be stable, since the output signal does not increase indefinitely in value.

The long-division method that we used to obtain the series given in Equation 4.28 is a valid procedure for evaluating algebraic operators generally. It is, however, tedious. Furthermore, we may be interested only in determining whether the operator is stable and not at all interested in evaluating r. An alternative approach for testing an operator's stability and also for easily evaluating its power series expansions is to expand the operator into *partial fractions.*

An algebraic rational fraction, consisting of the ratio of two polynomials, can be decomposed into the sum of terms in many different ways. By the ordinary long-division method just illustrated, it may be expanded into an infinite series. The summing of this infinite series poses difficulties that could be avoided if the series were finite. The method of partial fractions enables us to replace the original rational fraction by the sum of several *simpler* rational fractions each having denominators that are of lower degree than the original expression.

Consider, for instance, the rational expression

$$H(x) = \frac{x^3 + 4x^2 + 6x + 6}{x^2 + 3x + 2}. \tag{4.29a}$$

The denominator is of lower degree than the numerator so ordinary long division may be used to remove a polynomial in x:

$$\begin{array}{r} x + 1 \\ x^2 + 3x + 2 \overline{)\, x^3 + 4x^2 + 6x + 6} \\ \underline{x^3 + 3x^2 + 2x} \\ x^2 + 4x + 6 \\ \underline{x^2 + 3x + 2} \\ x + 4 \end{array}$$

Hence,

$$H(x) = x + 1 + \frac{x + 4}{x^2 + 3x + 2}. \tag{4.29b}$$

Next, we observe that the denominator of $H(x)$, which is also the denominator of the remainder term, has two zeros, at $x = -2$ and $x = -1$. Therefore, we may write the remainder $R(x)$ as

$$R(x) = \frac{x + 4}{(x + 1)(x + 2)} = \frac{k_1}{(x + 1)} + \frac{k_2}{(x + 2)} \tag{4.30a}$$

where the constants k_1 and k_2 must be determined to satisfy the equality.

A good method for determining, for example k_1, is to multiply Equation 4.30a through by the denominator $(x + 1)$ associated with k_1. This yields

$$(x + 1)R(x) = \frac{x + 4}{x + 2} = k_1 + (x + 1)\frac{k_2}{x + 2}. \tag{4.30b}$$

If we set $x = -1$ in this expression, the term containing k_2 vanishes, leaving only k_1 on the right-hand side. This expression may then be used to obtain the value of k_1, provided that we evaluate the left-hand side, which is indeterminant, $0 \cdot \infty$. This may be done either by canceling the zero factor from numerator and denominator or by using L'Hospital's rule. In any case,

$$\lim_{x \to -1} [(x + 1)R(x)] = 3 = k_1. \tag{4.30c}$$

Similarly, to obtain the value of k_2, multiply Equation 4.30a through by $(x + 2)$ and then set $x = -2$:

$$(x + 2)R(x) = \frac{x + 4}{x + 1} = (x + 2)\frac{k_1}{(x + 1)} + k_2$$

$$\lim_{x \to -2} [(x + 2)R(x)] = \frac{2}{-1} = -2 = k_2. \tag{4.30d}$$

Hence,

$$k_1 = 3 \qquad \text{and} \qquad k_2 = -2$$

and

$$\frac{x + 4}{(x + 1)(x + 2)} = \frac{3}{x + 1} - \frac{2}{x + 2}. \tag{4.30e}$$

In this way, we may decompose the complicated expression for $H(x)$ originally given by Equation 4.29a into a sum of several simpler expressions

$$\frac{x^3 + 4x^2 + 6x + 6}{x^2 + 3x + 2} = x + 1 + \frac{3}{x + 1} - \frac{2}{x + 2}. \tag{4.30f}$$

This operation of replacing a complicated expression by the sum of several less complicated expressions is very important. We shall use it again and again in following sections of this text.

For instance, suppose that in Equation 4.30f x were a linear, stationary operator. Then we could realize the operator H, after trivial manipulations on Equation 4.30f, by

$$H(x) = x + 1 + \frac{3}{1 + x} - \frac{1}{1 + \frac{1}{2}x}. \tag{4.30g}$$

Thus, as shown in the graph of Figure 4.69, the effect of the operation H on a signal f may be obtained by superimposing the effects of the three individual single-loop operators

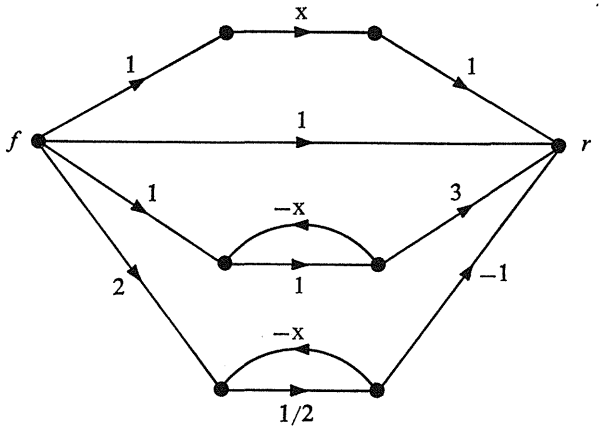


FIGURE 4.69

on the same signal f. This important tactic permits us to break a complicated operator, that is difficult to analyze, into the sum of simple operators which are individually much easier to treat.

To illustrate how the method of partial fractions may be used to determine if an operator is stable, suppose that we combine two of the single-delayor-loop operators "in parallel" as in Figure 4.70.

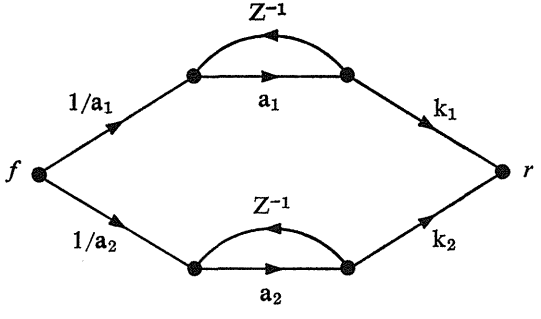


FIGURE 4.70 *An additive combination of two one-loop operators.*

Here,

$$H = \frac{k_1}{1 - a_1 Z^{-1}} + \frac{k_2}{1 - a_2 Z^{-1}}. \tag{4.31a}$$

Each of these terms is readily expanded into a "geometric" series:

$$H = k_1 \sum_{k=0,1,2,\ldots} (a_1)^k Z^{-k} + k_2 \sum_{k=0,1,2,\ldots} (a_2)^k Z^{-k}, \tag{4.31b}$$

and it is clear that the operator will be stable if

$$|a_1| < 1 \qquad \text{and} \qquad |a_2| < 1. \tag{4.31c}$$

If *either* of these two additive operators is unstable, it is clear that the overall system will be unstable. Furthermore, although we show only *two* additive operators, we could have used any number and reached the same conclusion: if any one of the additive single-loop operators is unstable, the entire system will be unstable.

Now, we have already shown that expressions such as Equation 4.31a, may be manipulated in accord with the familiar rules of algebra. In particular, we may combine the two additive terms into a single rational fraction:

$$H = \frac{k_1}{1 - a_1 Z^{-1}} + \frac{k_2}{1 - a_2 Z^{-1}},$$

$$H = \frac{k_1(1 - a_2 Z^{-1}) + k_2(1 - a_1 Z^{-1})}{(1 - a_1 Z^{-1})(1 - a_2 Z^{-1})}, \tag{4.32a}$$

$$H = \frac{(k_1 + k_2) - (k_1 a_2 + k_2 a_1) Z^{-1}}{1 - (a_1 + a_2) Z^{-1} + a_1 a_2 Z^{-2}}. \tag{4.32b}$$

But this form may be made identical to the operator shown in Figure 4.67:

$$H = \frac{Z^{-1}}{1 - 1.4Z^{-1} + 0.48Z^{-2}}$$

by setting

$$(k_1 + k_2) = 0, \tag{4.33a}$$

$$-(k_1 a_2 + k_2 a_1) = 1, \tag{4.33b}$$

$$a_1 + a_2 = 1.4, \tag{4.33c}$$

$$a_1 a_2 = 0.48. \tag{4.33d}$$

QUESTION 4.56 Solve Equations 4.33a–d and show that

$$a_1 = 0.8, \qquad k_1 = 5,$$

$$a_2 = 0.6, \qquad k_2 = -5.$$

In solving these equations, is there a preferred order?

Thus, we may expand the two-loop operator of Figure 4.66 by the method of partial fractions into an additive sum of two one-loop operators:

$$\frac{Z^{-1}}{1 - 1.4Z^{-1} + 0.48Z^{-2}} = \frac{5}{1 - 0.8Z^{-1}} - \frac{5}{1 - 0.6Z^{-1}}. \tag{4.34}$$

From this expansion, it is immediately clear that the two-loop operator is *stable.* Furthermore, the operator may be expanded into the sum of two geometric series:

$$\frac{Z^{-1}}{1 - 1.4Z^{-1} + 0.48Z^{-2}} = 5\sum_k (0.8)^k Z^{-k} - 5\sum_k (0.6)^k Z^{-k}$$

$$= 5\sum_{k=1,2,\ldots} [(0.8)^k - (0.6)^k] Z^{-k}. \tag{4.35}$$

This is easily done by forming two columns of numbers corresponding to the successive terms in two series, and then adding them to get the final scalar coefficients:

TABLE 4.1

k	$5(0.8)^k$	$-5(0.6)^k$	Coeff of Z^{-k}
0	5.00	−5.00	0.00
1	4.00	−3.00	1.00
2	3.20	−1.80	1.40
3	2.56	−1.08	1.48
4	2.05	−0.65	1.40
5	1.64	−0.39	1.25
6	1.31	−0.23	1.08
7	1.05	−0.14	0.91
8	0.84	−0.08	0.76
9	0.67	−0.05	0.62
10	0.54	−0.03	0.51
11	0.43	−0.02	0.41
12	0.34	−0.01	0.33

The numbers given in the right-hand column of Table 4.1 agree closely with the results previously obtained by the much harder (and less accurate) long-division method shown on page 287. Furthermore, the table clearly shows why for $t > 10$ the signal value decreased by 0.8 with each added time unit; beyond $k = 10$ the second geometric sequence is negligibly small compared to the first.

The general formula for an expansion, such as used in Equation 4.35, is easily obtained, once a partial fraction expansion of the operator has been obtained. Thus, if

$$H(x) = \sum_i \frac{c_i}{1 - a_i x},$$

by expanding each denominator in a power series in x,

$$\frac{1}{1 - a_i x} = 1 + a_i x + a_i^2 x^2 + \cdots$$

$$= \sum_{k=0}^{\infty} a_i^k x^k$$

and then collecting all terms involving x^k, we get

$$H(x) = \sum_k \left[\sum_i c_i a_i^k\right] x^k.$$

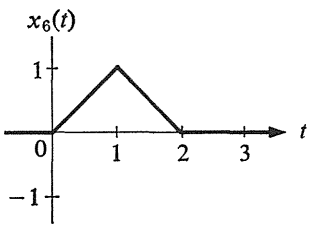
Equation 4.35 is of this form, with $a_1 = 0.8$, $a_2 = 0.6$, $c_1 = 5$, and $c_2 = -5$.

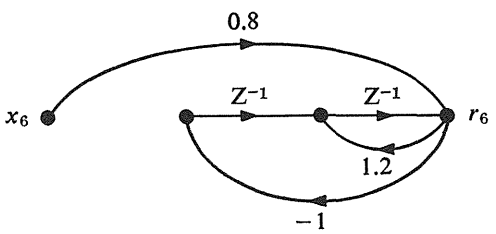
QUESTION 4.57 Find the first six terms in the power-series expansion of this operator, using the long-division method:

$$H = \frac{0.8}{1 - 1.2Z^{-1} + Z^{-2}}.$$

What difficulty do you encounter when you try to check your results by a partial-fraction expansion similar to that of Equation 4.35? (Answer)

QUESTION 4.58 The input to the operator shown below is the triangular signal x_6. Verify that the value of the output signal r_6 at the instants $t = 1, 2, 3, 4, 5, 6$, is equal to the value of the function sin (53.2 deg t) at these same instants. Plot $r_6(t)$ for $-1 < t < 6$. (Answer)



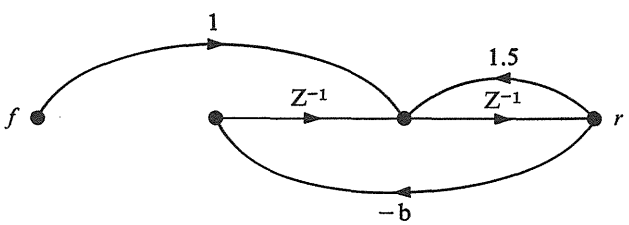


QUESTION 4.59 Is the operator

$$H = \frac{2.05Z^{-1}}{1 - 0.45Z^{-1} - Z^{-2}}$$

stable or unstable? When this operator is applied to the dot signal f_1, the signal f_1H is obtained. Use the partial-fraction method to find the value $f_1H(10.5)$ of the signal f_1H at the instant $t = 10.5$. (Answer)

QUESTION 4.60 Consider the operator graph shown.



When the transmittance $-b$ of the outer feedback loop is zero, the operator is clearly unstable because it reduces to a one-loop configuration with $a = 1.5$.

1. Make a partial-fraction expansion of the graph operator by factoring the graph determinant into the product of two factors as in Equation 4.32a:

$$(1 - a_1Z^{-1})(1 - a_2Z^{-1}),$$

and expressing a_1 and a_2 in terms of the transmittance $-b$.

2. For stability, it is necessary that $|a_1| < 1$ and $|a_2| < 1$. Use these two conditions to determine the largest and smallest values between which b must lie if the operator is to be stable. (Answer)

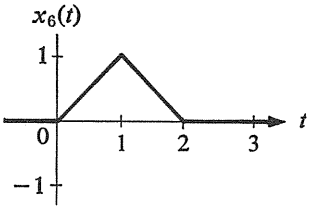
ANSWER TO QUESTION 4.57

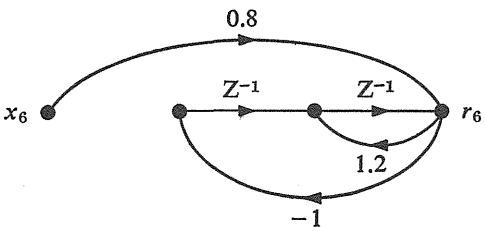
By ordinary long division, one obtains

$$\frac{0.8}{1 - 1.2Z^{-1} + Z^{-2}} = 0.8 + 0.96Z^{-1} + 0.352Z^{-2} - 0.538Z^{-3} - 1.00Z^{-4} - 0.661Z^{-5} + \cdots.$$

ANSWER TO QUESTION 4.58

The graph operator between x_6 and r_6





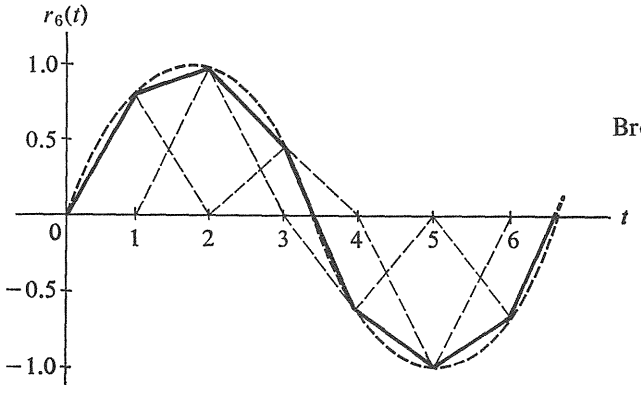
is precisely the operator considered in Question 4.57. Using the expansion found there, we may express r_6 in terms of x_6 as

$$r_6 = x_6[0.8 + 0.96Z^{-1} + 0.352Z^{-2} - 0.538Z^{-3} - 1.000Z^{-4} - 0.661Z^{-5} + \cdots],$$

as illustrated by the graph. In fact, at integer values of t, the value of r_6 is given by

$$r_6(n) = \sin(53.2n \text{ deg}).$$

See also Figure 6.4.



Broken curve: sin (53.2t deg)

ANSWER TO QUESTION 4.59 To investigate the stability of any operator, we need only investigate the graph determinant (which is the part of the algebraic expression that describes the loops). By first factoring the graph determinant, we may replace the original operator by the sum of one-loop operators using the partial-fraction method:

$$\mathrm{H} = \frac{2.05\mathrm{Z}^{-1}}{1 - 0.45\mathrm{Z}^{-1} - \mathrm{Z}^{-2}} = \frac{k_1}{1 + 0.80\mathrm{Z}^{-1}} + \frac{k_2}{1 - 1.25\mathrm{Z}^{-1}}.$$

To find k_1 and k_2, we equate numerator expressions

$$2.05\mathrm{Z}^{-1} = k_1(1 - 1.25\mathrm{Z}^{-1}) + k_2(1 + 0.8\mathrm{Z}^{-1}),$$

where $k_1 = -1$ and $k_2 = +1$.

The second one-loop operator, for which $a_2 = 1.25$, is clearly *unstable* and will produce a growing response, whereas the first one-loop operator produces a decaying response:

$$\mathrm{H} = \sum_{k=0,1,2,\ldots} [-(-0.8)^k + (1.25)^k]\mathrm{Z}^{-k}.$$

For $k = 10$, the coefficient is

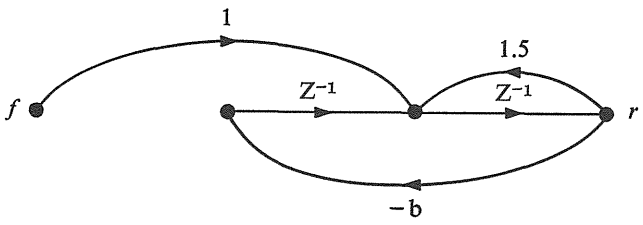
$$-(-0.8)^{10} + (1.25)^{10} = -0.106 + 9.30$$
$$= \mathbf{9.194.}$$

Hence, for a dot f_1 input, the response signal value at $t = 10.5$ is

$$f_1[9.194\mathrm{Z}^{-10}](10.5) = 9.194,$$

since all the other terms are zero.

ANSWER TO QUESTION 4.60



1. Here $H = \dfrac{Z^{-1}}{1 - 1.5Z^{-1} + bZ^{-2}}$. By comparison with Equation 4.32b,

$$a_1 + a_2 = 1.5,$$
$$a_1 a_2 = b.$$

Hence,

$$a_1^2 + a_1 a_2 = 1.5a_1$$

or

$$a_1^2 - 1.5a_1 + b = 0.$$

Thus,

$$a_1 = \tfrac{3}{4} + \sqrt{(\tfrac{3}{4})^2 - b},$$
$$a_2 = \tfrac{3}{4} - \sqrt{(\tfrac{3}{4})^2 - b}.$$

Then,

$$\frac{Z^{-1}}{1 - 1.5Z^{-1} + bZ^{-2}} = \frac{1}{a_1 - a_2}\left[\frac{1}{1 - a_1 Z^{-1}} - \frac{1}{1 - a_2 Z^{-1}}\right].$$

2. We wish to determine the range of values of b for which $|a_1| < 1$ and $|a_2| < 1$. Evidently, for $|a_1| < 1$ it is necessary that

$$\sqrt{(\tfrac{3}{4})^2 - b} < 0.25.$$

Squaring both sides of the inequality yields

$$0.5625 - b < 0.0625.$$

Hence,

$$\mathbf{0.5 < b.}$$

For $b > (\frac{3}{4})^2$, the radical becomes imaginary,

$$a_1 = \tfrac{3}{4} + j\sqrt{b - (\tfrac{3}{4})^2}$$
$$a_2 = \tfrac{3}{4} - j\sqrt{b - (\tfrac{3}{4})^2} \qquad (j^2 = -1).$$

The *magnitude* of a complex number is given by the square root of the sum of the squares of its real and its imaginary parts:

$$|a_1| = [(\tfrac{3}{4})^2 + b - (\tfrac{3}{4})^2]^{1/2} = \sqrt{b}.$$

Also,

$$|a_2| = \sqrt{b}.$$

Hence,

$$b < 1.$$

The allowable range of b for which the operator is stable is therefore

$$0.5 < b < 1.$$

Summary

1. Operators may be classified according to three major properties:
   a) An operator is *static* if the value of the output signal at any instant depends *only* on the value of the input signal at that same instant. Otherwise, the operator is *dynamic.*
   b) An operator is *linear* if: (1) it commutes with a scalor and (2) the principle of superposition applies. Otherwise, the operator is *nonlinear.*
   c) An operator is *stationary* if it commutes with any delayor. Otherwise, it is *time-varying.*
2. Linearity is a sufficient property to permit flow-graph reductions, but Mason's rule may be used (in its simple form) only when the operators *commute.* This requires that the operators be *stationary* as well as *linear.*
3. A delayor Z^{-T} yields an output signal whose value at any instant, t, is equal to the value of the input signal at the earlier instant, $t - T$. Hence, a delayor is the basic *dynamic* operator. Using scalors and delayors, we may generate new signals of unlimited variety from a single primitive signal, such as a "dot."
4. The presence of loops involving delayors means that signals may be operated upon infinitely many times as they flow through an operator graph. If for *some* input signal of bounded value, the output signal increases without bound, the operator is said to be *unstable.* An operator is stable if the value of the output signal ultimately approaches zero subsequent to reducing all input signal values permanently to zero.
5. A common method for investigating stability of an operator is to decompose it into an additive combination of simpler operators each of whose stability can be more easily tested. Partial-fraction and power-series expansions are useful techniques for doing this.

INTRODUCTORY SYSTEMS AND DESIGN

W. H. Huggins | *Doris R. Entwisle*

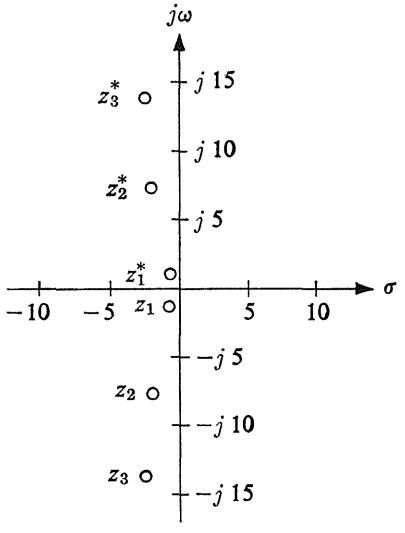
THE JOHNS HOPKINS UNIVERSITY

BLAISDELL PUBLISHING COMPANY

A DIVISION OF GINN AND COMPANY

WALTHAM, MASSACHUSETTS • TORONTO • LONDON

This process may be continued indefinitely, producing an infinite array of zeros of which the smallest six are shown in Figure 9.3.



Zeros of $s + e^{-s}$

$z_1 = -0.316 - j1.338,$

$z_2 = -2.065 - j7.58,$

$z_3 = -2.65 - j13.92,$

etc.

FIGURE 9.3 *The first six zeros of* $H(s) = s + e^{-s}$.

Iterative Numerical Computations

In finding the zeros of the function $H(s) = s + e^{-s}$, we used a method of successive approximations. This is an example of an iterative procedure that is particularly well suited to computers. Only for the simplest functions is it possible to write explicit closed-form solutions for their zeros. For simple polynomials, closed-form expressions for the zeros can be obtained only if the polynomial is of a degree less than 5; and even for polynomials of higher degree than quadratic you will find the iterative methods next discussed to be of indispensable value.

Newton's method • The simplest iterative scheme for finding zeros is Newton's method. Consider the arbitrary function $F(\cdot)$, which varies smoothly in the vicinity of its zeros. Then, it is possible to refine a good initial guess by using the Taylor's series expansion of the function about the point $s + \Delta s$:

$$F(s) = F(s + \Delta s) - F'(s + \Delta s)\Delta s + F''(s + \Delta s)\frac{(\Delta s)^2}{2!} - \cdots. \tag{9.7}$$

Suppose that s is a zero of $F(\cdot)$, so that $F(s) = 0$, and our initial guess, $s + \Delta s$, is in error by Δs. Then, if $F''(s + \Delta s)$ and Δs are sufficiently small, the series may often be approximated well by only the first two terms:

$$0 \doteq F(s + \Delta s) - F'(s + \Delta s)\Delta s.$$

Hence,

$$\Delta s \doteq \frac{F(s + \Delta s)}{F'(s + \Delta s)} = \Delta \hat{s}. \tag{9.8}$$

That is, by dividing the value of the function by the value of the derivative of the function both evaluated at $s + \Delta s$, we obtain an *estimate* $\Delta \hat{s}$ of the error Δs in our initial guess. By subtracting $\Delta \hat{s}$ from the original guess, a more accurate estimate, $s + \Delta s - \Delta \hat{s}$, of the zero is obtained. This process is repeated until $F(\cdot)$ is negligibly small and there is no further change in the estimate of the zero.

Just as Figure 9.2 illustrates the iterative method, so we may describe Newton's method by the signal generator shown in Figure 9.4. In the process of Figure 9.4, the

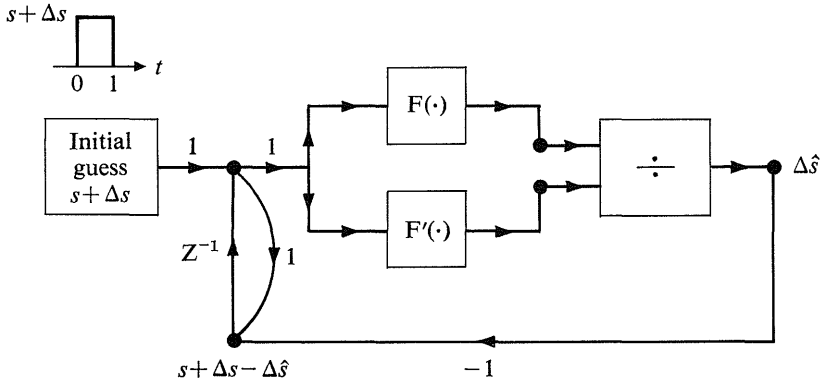


FIGURE 9.4 *Operational description of Newton's method for evaluating a zero of $F(s)$.*

function and its derivative are represented by the static nonlinear operators $F(\cdot)$ and $F'(\cdot)$ respectively. The initial guess is again represented by a dot signal of unit duration having an amplitude equal to the value $s + \Delta s$ of the initial guess. The output from the divider is a dot signal whose amplitude $\Delta \hat{s}$ is an estimate of the error Δs in the initial guess. By subtracting $\Delta \hat{s}$ from the initial guess, we obtain an improved estimate $s + \Delta s - \Delta \hat{s}$ which by means of a unit delayor may be used as an improved guess during the second time interval, etc.

To illustrate Newton's method, let us find the real value of s for which the transmittance function $F(s) = e^{-s} - s$ is zero. In applying this method, the better the initial estimate of the zero, the more quickly the process will converge to the final answer. It is therefore desirable to sketch the behavior of the function as in Figure 9.5. As shown in Figure 9.5, $F(\sigma)$ has a single zero occurring approximately at $\sigma = 0.5$. Hence, we use this as the initial guess and apply Newton's method. Since

$$F(\sigma) = e^{-\sigma} - \sigma,$$

$$F'(\sigma) = -e^{-\sigma} - 1,$$

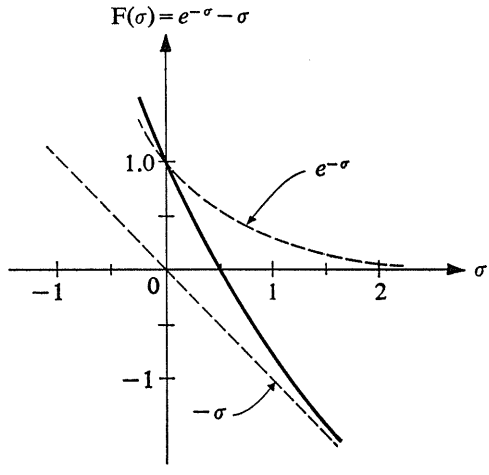


FIGURE 9.5 *A plot of* $F(\sigma) = e^{-\sigma} - \sigma$.

for $\sigma = +0.5$, we find

$$\Delta\hat{s} = \frac{F(0.5)}{F'(0.5)} = \frac{0.106}{-1.606} = -0.0606$$

and the improved estimate of the zero is $0.5 - (-0.0606) = 0.5606$. The new estimate of the error is then

$$\Delta\hat{s} = \frac{F(0.5606)}{F'(0.5606)} = \frac{0.009}{-1.570} = -0.0057,$$

and hence, within slide-rule accuracy, the zero of $F(\sigma)$ is given by

$$0.5606 + 0.0057 = \mathbf{0.566.}$$

QUESTION 9.5 Given the polynomial

$$F(s) = s^5 + 40,$$

use Newton's method to find the *real* zero of $F(s)$. (Answer)

Newton's method can also be used to refine an estimate of a *complex zero* as well as a real zero. For instance, let us find a complex zero of $F(s) = s^2 + s + 1$ by choosing $s = j1$ as the initial guess. Then

$$\Delta\hat{s} = \frac{F(j1)}{F'(j1)} = \frac{j1}{1 + j2} = 0.4 + j0.2.$$

Hence, a better estimate of the zero is

$$(0 + j1) - (0.4 + j0.2) = -0.4 + j0.8.$$

Repeating the calculation using this new estimate, we find

$$\Delta\hat{s} + \frac{F(-0.4 + j0.8)}{F'(-0.4 + j0.8)} = \frac{0.12 + j0.16}{0.2 + j1.6} = 0.106 - j0.0612.$$

Hence, a further estimate of the zero is

$$(-0.4 + j0.2) - (0.106 - j0.0612) = -0.506 + j0.8612.$$

Clearly, this sequence of values is rapidly approaching the "true" value $-\frac{1}{2} + j\sqrt{3}/2$ obtained by using the quadratic formula.

Incidentally, you will find it useful to use the remainder theorem to evaluate a "real" polynomial when the argument is complex, as illustrated by Questions 7.21 and 7.22 of Chapter 7. Thus, to evaluate $F(s)$ for $s = -0.4 + j0.8$, we first construct the "real" quadratic divisor

$$D(s) = s^2 + 0.8s + 0.6,$$

which vanishes for $s = -0.4 + j0.8$. Then dividing $F(s)$ by $D(s)$, yields a remainder polynomial $R(s)$,

$$\begin{array}{r} 1 \\ s^2 + 0.8s + 0.8 \overline{)\, s^2 + \quad s + \quad 1} \\ \underline{s^2 + 0.8s + 0.8} \\ \mathbf{0.2}s + \mathbf{0.2} = R(s). \end{array}$$

Hence,

$$\begin{aligned} F(-0.4 + j0.8) &= R(-0.4 + j0.8) \\ &= 0.2(-0.4 + j0.8) + 0.2 \\ &= \mathbf{0.12 + j0.16.} \end{aligned}$$

The advantage of using the remainder theorem is rather insignificant in this example because $F(s)$ is only of the second degree. However, consider Question 9.6.

QUESTION 9.6 Find *all* the zeros of the polynomial

$$P(s) = s^4 + 3s^3 + 5s^2 + 4s + 3.$$

As an initial guess, try $s = -0.5 + j1$ and refine your answer, using Newton's method and the remainder theorem to evaluate $P(s)/P'(s) = \Delta\hat{s}$. Correct the original guess and repeat the process until $\Delta\hat{s} = 0$. (Four iterations are required to obtain the zero to within slide-rule accuracy, the successive approximations being

Initial guess: $-0.500 + j1.000$,

First iteration: $-0.085 + j0.889$,

Second iteration: $-0.221 + j0.945$,

Third iteration: $-0.227 + j1.015$,

Fourth iteration: $-0.225 + j1.014$).

(Answer)

Stability and convergence • Before continuing our study of zeros and poles, you will find it worthwhile to examine why an iterative numerical procedure, such as that shown in Figure 9.2, converges to the correct answer. Iterative methods for improving an initial guess do not always converge—they may sometimes become unstable, and an initial error is magnified rather than diminished by each iteration.

ANSWER TO QUESTION 9.5 Examining the set of coefficients of $F(s) = s^5 + 40$, one finds *no changes* of sign. Hence, by Descartes' Rule of Signs, there are *no positive* real zeros. When s is replaced by $-s$, the coefficients exhibit *one* change in sign. Hence, there is *at most, one negative* real zero. The remaining four zeros must be complex.

As a crude initial guess of the negative zero, we consider $s = -2$. Then

$$F(s) = s^5 + 40, \qquad F'(s) = 5s^4,$$
$$F(-2) = 8, \qquad F'(-2) = 80.$$

Hence, by Equation 9.8, the error in our initial estimate is approximately

$$\Delta\hat{s} = \frac{F(-2)}{F'(-2)} = \frac{8}{80} = 0.1,$$

giving as a closer estimate, $-2 - 0.1 = -2.1$. Repeating the calculation for $s = -2.1$, we obtain

$$\Delta\hat{s} = \frac{F(-2.1)}{F'(-2.1)} = \frac{-0.9}{97.5} = -0.0092,$$

so a good estimate of the negative real zero is

$$s = -2.1 + 0.0092 = \mathbf{-2.091.}$$

To verify that this is indeed a zero of $F(s)$, we may show that $(s + 2.091)$ is a factor of $F(s)$:

$$\begin{array}{r|l}
 & s^4 - 2.091s^3 + 4.37s^2 - 9.15s + 19.15 \\
\hline
s + 2.091 & s^5 + 0s^4 + 0s^3 + 0s^2 + 0s + 40 \\
 & \underline{s^5 + 2.091s^4} \\
 & \quad -2.091s^4 + 0s^3 \\
 & \quad \underline{-2.091s^4 - 4.37s^3} \\
 & \qquad\qquad 4.37s^3 + 0s^2 \\
 & \qquad\qquad \underline{4.37s^3 + 9.15s^2} \\
 & \qquad\qquad\quad -9.15s^2 + 0s \\
 & \qquad\qquad\quad \underline{-9.15s^2 - 19.15s} \\
 & \qquad\qquad\qquad\quad +19.15s + 40 \\
 & \qquad\qquad\qquad\quad \underline{19.15s + 40} \\
 & \qquad\qquad\qquad\qquad\quad 0s + 0
\end{array}$$

The remainder vanishes, verifying that $(s + 2.091)(s^4 - 2.091s^3 + 4.37s^2 - 9.15s + 19.15)$ are factors of $F(s)$. The zeros of the fourth degree factor are all complex, as noted previously.

To investigate the stability of the process, one usually makes a perturbation analysis of the numerical system. Thus, with reference to Figure 9.2, suppose that we denote our original guess by $\varphi + \Delta\varphi$, where φ now represents the true value (i.e., 76.7 deg or 1.338 rad for the first zero), and $\Delta\varphi$ is the amount by which our guess is in error. Then, likewise, $\hat{r} = r + \Delta r$ and $\cos\hat{\varphi} = \cos\varphi + \Delta(\cos\hat{\varphi})$ and $\hat{\varphi} = \varphi + \Delta\hat{\varphi}$ where each value is expressed as a *perturbation* about its *true* value. The reason for doing this is that for sufficiently small perturbations, many nonlinear static operators may be replaced by *equivalent linear* transmittances that describe approximately the relations among these small perturbations. (The transmittance is given by the *slope* of the input–output function at the operating point.) In other words, we can replace the grossly *nonlinear relations* of Equation 9.6a–c by an approximate set of *linear relations* among the perturbations.

ANSWER TO QUESTION 9.6 To refine the initial guess, we must evaluate $P(s)/P'(s)$ for $s = -0.5 + j1.0$. To use the remainder theorem, we first form a real quadratic polynomial $D(s)$ which is zero for $s = -0.5 + j1.0$. By the method discussed in Question 7.19 of Chapter 7, this polynomial is found to be

$$D(s) = s^2 + s + 1.25.$$

Then, dividing $D(s)$ into $P(s)$, we obtain

$$\begin{array}{rrrl|llll} & & & & 1 & 2 & 1.75 & \\ \hline 1 & 1 & 1.25 &) & 1 & 3 & 5 & 4 \quad 3 \\ & & & & 1 & 1 & 1.25 & \\ \hline & & & & & 2 & 3.75 & 4 \\ & & & & 2 & 2 & & 2.5 \\ \hline & & & & & 1.75 & 1.5 & 3 \\ & & & & & 1.75 & 1.75 & 2.19 \\ \hline & & & & & & \mathbf{-0.25} & \mathbf{0.81} \end{array}$$

Hence,

$$R(s) = -0.25s + 0.81.$$

Evaluating $R(s)$ for $s = -0.5 + j1.0$, we find that

$$\begin{aligned} P(-0.5 + j1) &= -0.25(-0.5 + j1) + 0.81 \\ &= \mathbf{0.935 - j0.25.} \end{aligned}$$

In a similar fashion, we may evaluate $P'(s) = 4s^3 + 9s^2 + 10s + 4$ for this same value of $s = -0.5 + j1.0$, by dividing $P'(s)$ by $D(s)$ and evaluating the remainder:

$$\begin{array}{rrrl|llll} & & & & 4 & 5 & & \\ \hline 1 & 1 & 1.25 &) & 4 & 9 & 10 & 4 \\ & & & & 4 & 4 & 5 & \\ \hline & & & & & 5 & 5 & 4 \\ & & & & & 5 & 5 & 6.25 \\ \hline & & & & & & \mathbf{0} & \mathbf{-2.25.} \end{array}$$

Here,

$$R(s) = -2.25.$$

Hence,

$$\left.\frac{P(s)}{P'(s)}\right|_{s=-0.5+j1} = \frac{0.935 - j0.25}{-2.25} = \mathbf{-0.415 + j0.111.}$$

A more accurate estimate of the zero is obtained by subtracting $\Delta\hat{s} = P(s)/P'(s)$ from the previous estimate:

Initial guess:	$-0.5 \quad + j1$
$-\Delta\hat{s}$:	$+0.415 - j0.111$
First iteration:	$\mathbf{-0.085 + j0.889.}$

The computation is now repeated using $s = -0.085 + j0.889$, for which the quadratic divisor is

$$D(s) = s^2 + 0.17s + 0.862.$$

This yields the error estimate

$$\Delta\hat{s} = 0.136 - j0.0572,$$

and the second estimate of the zero is now $-0.221 + j0.945$. At the fourth iteration, the quadratic divisor is

$$D(s) = s^2 + 0.45s + 1.079$$

and the remainder vanishes:

$$\begin{array}{rrrrrrr}
 & & & 1 & 2.55 & 2.785 & \\
 & & & \multicolumn{4}{l}{\text{---------------------}} \\
1 & 0.45 & 1.079 & 1 & 3 & 5 & 4 \quad 3 \\
 & & & 1 & .45 & 1.079 & \\
 & & & \multicolumn{4}{l}{\text{---------------------}} \\
 & & & & 2.55 & 3.921 & 4 \\
 & & & & 2.55 & 1.136 & 2.75 \\
 & & & & & \multicolumn{2}{l}{\text{---------------}} \\
 & & & & & 2.785 & 1.25 \quad 3 \\
 & & & & & 2.785 & 1.25 \quad 3 \\
 & & & & & & \multicolumn{2}{l}{\text{---------------}} \\
 & & & & & & \mathbf{0} \quad \mathbf{0.}
\end{array}$$

Hence, the quadratic divisor is a factor of $P(s)$, the remaining factor being the quotient polynomial

$$P(s) = (s^2 + 0.45s + 1.079)(s^2 + 2.55s + 2.785).$$

In this case, the quotient is also a quadratic polynomial, and the zeros of both quadratic factors may be immediately written by use of the quadratic formula:

$$z_{1,2} = -0.225 \pm j1.014,$$
$$z_{3,4} = -1.275 \pm j1.073.$$

If the original polynomial had been, for example, of degree 6, the quotient obtained in the last step after dividing out $D(s)$ would have been of degree 4. This same process could then be applied to extract another quadratic factor from this quotient polynomial. In this way, a real polynomial of arbitrarily high degree may ultimately be reduced to the product of real quadratic factors, from which the individual zeros are easily calculated.

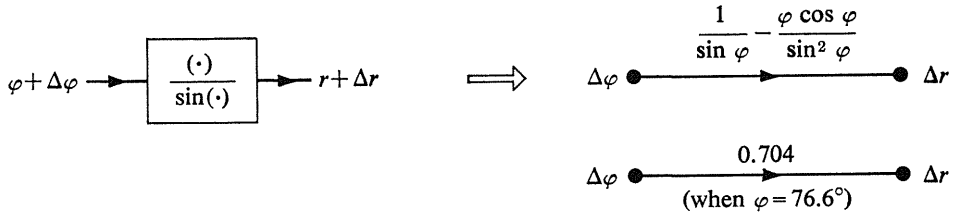
For instance, consider the function, $\varphi/\sin\varphi$. If φ is increased to $\varphi + \Delta\varphi$, the new value of the function may be expressed by a Taylor's series expansion:

$$\frac{\varphi + \Delta\varphi}{\sin(\varphi + \Delta\varphi)} = \frac{\varphi}{\sin\varphi} + \frac{d}{d\varphi}\left[\frac{\varphi}{\sin\varphi}\right]\Delta\varphi + \cdots.$$

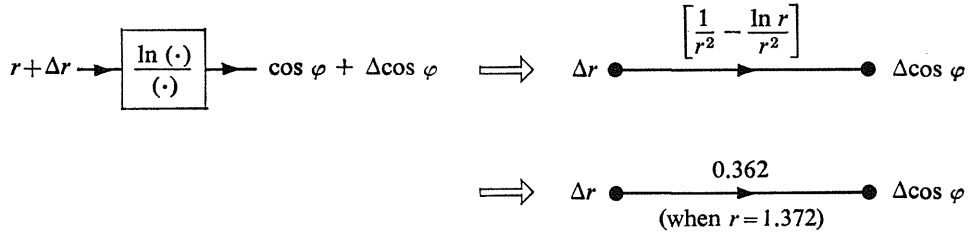
That is, the *change* in the value of the function is given to first-order terms in $\Delta\varphi$ by

$$\Delta r \doteq \frac{d}{d\varphi}\left[\frac{\varphi}{\sin\varphi}\right]\Delta\varphi = \left[\frac{1}{\sin\varphi} - \frac{\varphi\cos\varphi}{\sin^2\varphi}\right]\Delta\varphi.$$

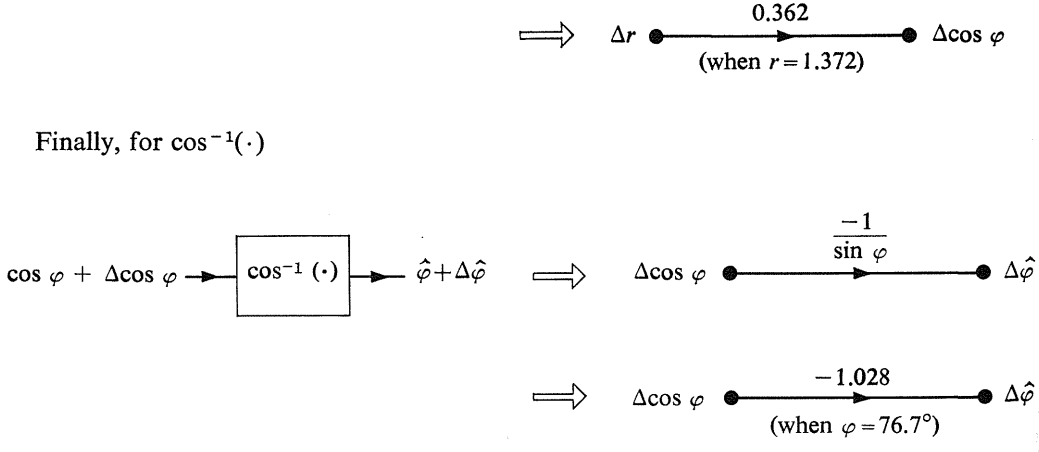
At the true solution, where $\varphi = 76.7$ deg or 1.338 rad, the value of the derivative is 0.704. For small changes in φ about the point 76.7 deg, the nonlinear operator $\varphi/\sin\varphi$ acts like a scalor of transmittance, 0.704:



Similarly, by differentiating the function $\ln(\cdot)/(\cdot)$ with respect to its argument $(\cdot)$, we find



Finally, for $\cos^{-1}(\cdot)$



Hence, the small error $\Delta\varphi$ in the initial guess will result in an error $\Delta\hat{\varphi}$ in $\hat{\varphi}$, as shown by Figure 9.6. The error $\Delta\hat{\varphi}$ in $\hat{\varphi}$ is therefore of *opposite sign* and only about 27% as *large*.

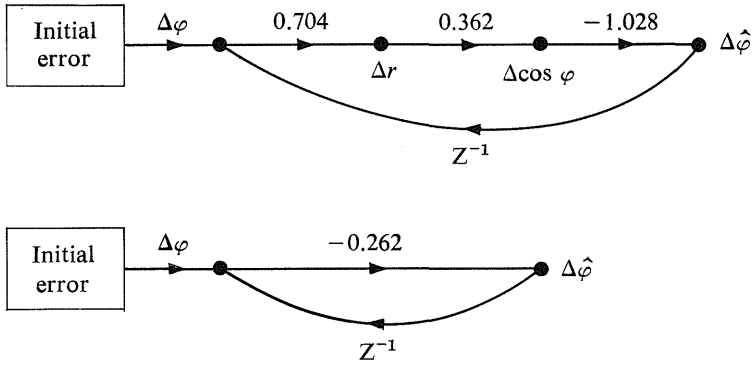


FIGURE 9.6 *A linearized graph of the process of Figure 9.2 at its first zero.*

Thus, $\hat{\varphi}$ is a *better* estimate of the true value than our original guess (provided $\Delta\varphi$ is sufficiently small to begin with). This is confirmed by the calculations given in Table 9.1, where for our initial guess of $\varphi = 80$ deg the error is

$$\Delta\varphi = 80 \text{ deg} - 76.7 \text{ deg} = 3.3 \text{ deg}.$$

The error in $\hat{\varphi}$ is seen to be

$$\Delta\hat{\varphi} = 75.7 \text{ deg} - 76.7 \text{ deg} = -1.0 \text{ deg},$$

which is of *opposite sign* and about 30% of the original error. In fact a plot of the successive errors, illustrated by Figure 9.7, shows that they alternate in sign as they decay to

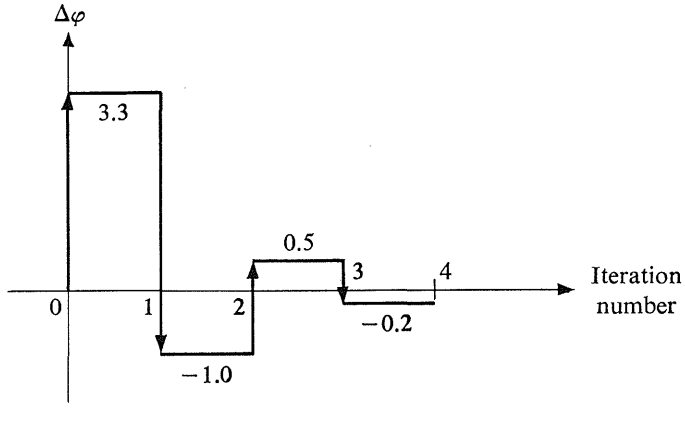


FIGURE 9.7 *The error $\Delta\varphi$ converges to zero on successive iterations.*

zero. As discussed in Chapter 4, this sequence of operations is *stable* because the loop transmittance is smaller than 1. Hence, the errors die away and the answer converges to the true value. If you make a perturbation analysis for the second zero, where $\varphi = 74.75$ deg and $r = -7.85$, you will find that the loop transmittance is smaller yet, thus leading to still more rapid convergence to the correct answer.

This illustrates a general principle in iterative numerical calculations: one should always arrange one's successive calculations so that the errors die away rather than grow. Each specific case may be investigated using the perturbation method just described.

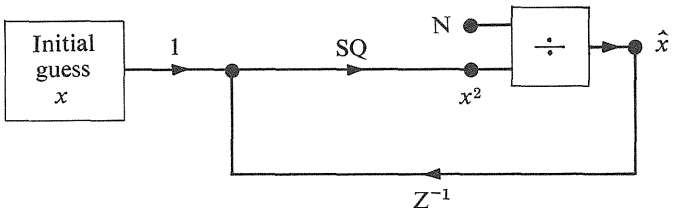
QUESTION 9.7 You wish to calculate the real zero of the polynomial $x^3 - N$, where N is a constant. Evidently, x is the cube root of N. Furthermore, any of the following relations apply:

$$x^3 = N$$

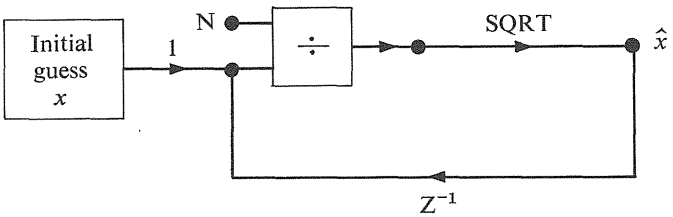
$$x^2 = \frac{N}{x}$$

$$x = \frac{N}{x^2}$$

You, therefore, may construct either of the iterative schemes shown below, although only one will converge to the correct answer for all N.

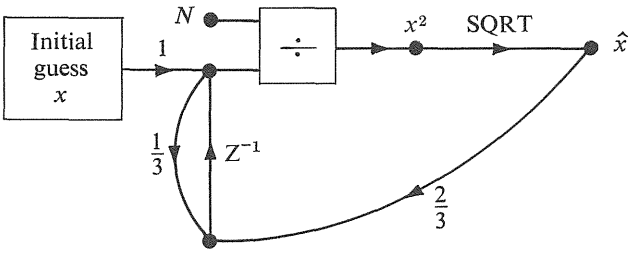


or



Can you perform perturbation analyses for each of these two schemes to decide which one will converge to the cube root of N? (Answer)

QUESTION 9.8 Another possibility for finding the cube root of N is shown here.



Use it to evaluate $\sqrt[3]{48}$, starting with $x = 3$ as a (poor) initial guess.

Graphical Interpretation of the Zeros

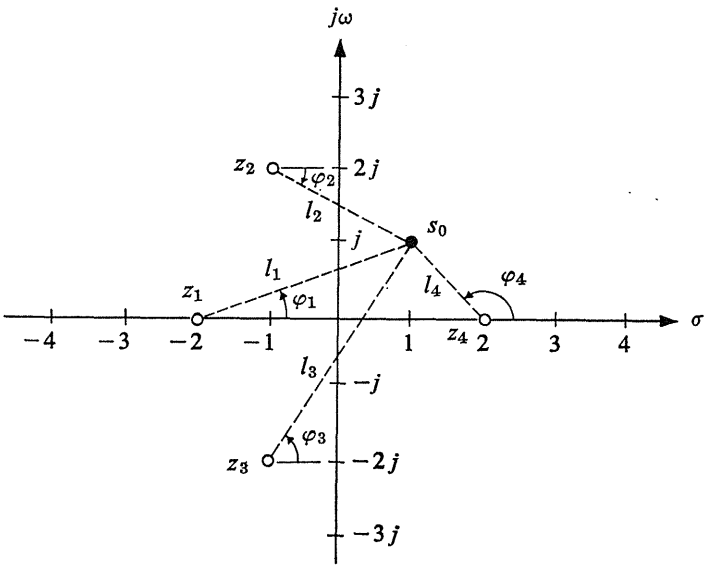
When plotted on the s-plane, the zeros of a transmittance function provide very clear and useful information about the behavior of that function for other values of s as well. This is particularly true when the transmittance function is an ordinary nth degree polynomial:

$$P(s) = a_0 s^n + a_1 s^{n-1} + \cdots + a_{n-1} s + a_n. \tag{9.9a}$$

Then the polynomial may be expressed exactly as a product of a scale factor and n first-degree factors:

$$P(s) = a_0(s - z_1)(s - z_2) \ldots (s - z_{n-1})(s - z_n). \tag{9.9b}$$

Both ways of expressing $P(s)$ involve $n + 1$ parameters, and either may be used to find the value of $P(s)$ for any s. However, Equation 9.9b very clearly shows that if the value of s is *close* to any of the zeros $z_1, z_2, \ldots, z_n$, then the value of $P(s)$ will likely be small, whereas if s is far away from all the zeros, $P(s)$ will have a large magnitude. Hence, the location of the zeros in the s-plane offers a valuable pictorial aid for understanding the behavior of the function for all values of its argument.



$z_1 = -2 + j0,$
$z_2 = -1 + j2,$
$z_3 = -1 + j2,$
$z_4 = \ \ 2 + j0.$

FIGURE 9.8 *The evaluation of a polynomial at* $s = s_0$.

In Figure 9.8, we show the zeros of a certain polynomial and also the value of the scale factor ($SF = 1.0$). These data completely determine the polynomial $P(s)$. We also show a geometrical construction for evaluating P at any point s_0 in the complex frequency plane. At s_0, Equation 9.9b may be written:

$$P(s_0) = a_0(s_0 - z_1)(s_0 - z_2) \ldots (s_0 - z_{n-1})(s_0 - z_n). \tag{9.9c}$$

Each of the factors in this equation has a simple geometrical meaning: it is the "complex distance" from the zero to the point s_0. Thus, the first factor $(s_0 - z_1)$ is the complex distance from z_1 to s_0. It is shown in Figure 9.8 as a broken line of length l_1 and angle

This illustrates a general principle in iterative numerical calculations: one should always arrange one's successive calculations so that the errors die away rather than grow. Each specific case may be investigated using the perturbation method just described.

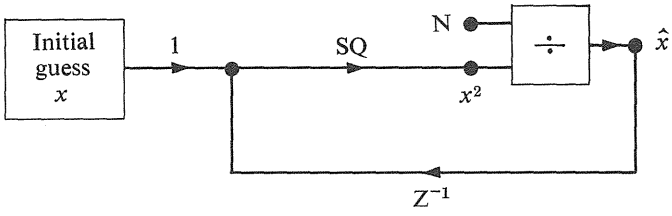
QUESTION 9.7 You wish to calculate the real zero of the polynomial $x^3 - N$, where N is a constant. Evidently, x is the cube root of N. Furthermore, any of the following relations apply:

$$x^3 = N$$

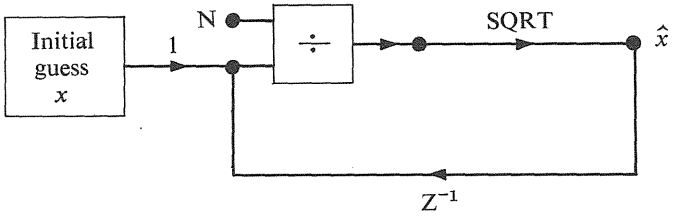
$$x^2 = \frac{N}{x}$$

$$x = \frac{N}{x^2}$$

You, therefore, may construct either of the iterative schemes shown below, although only one will converge to the correct answer for all N.

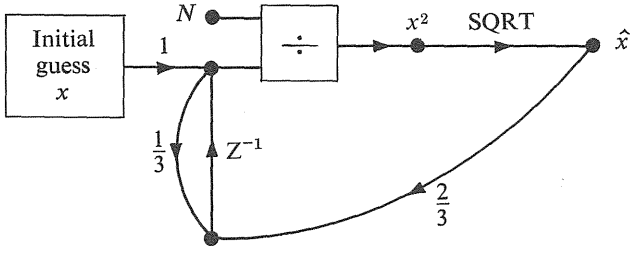


or



Can you perform perturbation analyses for each of these two schemes to decide which one will converge to the cube root of N? (Answer)

QUESTION 9.8 Another possibility for finding the cube root of N is shown here.



Use it to evaluate $\sqrt[3]{48}$, starting with $x = 3$ as a (poor) initial guess.

INTRODUCTORY SYSTEMS AND DESIGN

W. H. Huggins | *Doris R. Entwisle*

THE JOHNS HOPKINS UNIVERSITY

BLAISDELL PUBLISHING COMPANY

A DIVISION OF GINN AND COMPANY

WALTHAM, MASSACHUSETTS • TORONTO • LONDON

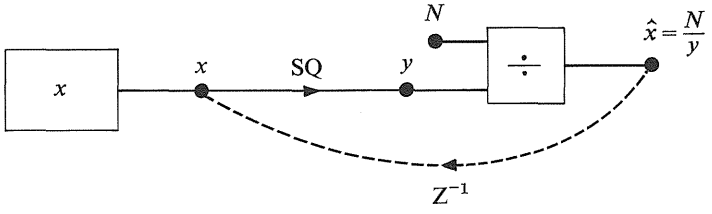
φ_1 which you will note is simply the polar representation for $s_0 - z_1$, that is, $(s_0 - z_1) = l_1\underline{/\varphi_1}$. Likewise, each of the other factors corresponds exactly to the complex distance from its zero to the point s_0, and in polar form may be written as $(s_0 - z_2) = l_2\underline{/\varphi_2}$, $(s_0 - z_3) = l_3\underline{/\varphi_3}$, and $(s_0 - z_4) = l_4\underline{/\varphi_4}$. By Equation 9.9b, the value of the polynomial is then expressed as

$$\begin{aligned} P(s_0) &= a_0(l_1\underline{/\varphi_1})(l_2\underline{/\varphi_2})(l_3\underline{/\varphi_3})(l_4\underline{/\varphi_4}) \\ &= a_0 l_1 l_2 l_3 l_4\underline{/\varphi_1 + \varphi_2 + \varphi_3 + \varphi_4}. \end{aligned} \tag{9.10}$$

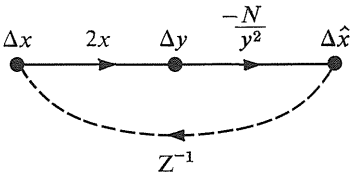
Hence, at $s = s_0$ the polynomial has a complex value whose *magnitude* is the product of the *magnitudes* $l_1, l_2, \ldots$ and whose *angle* is the *sum* of the *angles* $\varphi_1, \varphi_2, \ldots$. From these simple geometrical relations, it is easy to visualize how the magnitude and angle of $P(s_0)$ will vary as the point s_0 is moved around in the s-plane.

ANSWER TO QUESTION 9.7

Consider the first scheme:



At a true solution, $y = x^2$, $N = x^3$, and $\hat{x} = x$. However, if x is changed by a small amount, Δx, the consequent changes in y and $\hat{x}$ are given by the transmittances as shown.



These transmittances are obtained by differentiating the input–output function with respect to its input variable. Thus,

$$x^2 = y,$$

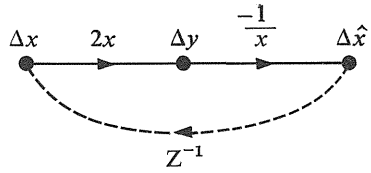
$$2x\Delta x = \Delta y,$$

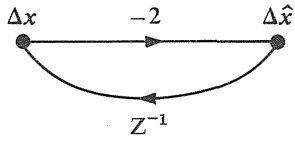
and

$$\frac{N}{y} = \hat{x},$$

$$-\frac{N}{y^2}\Delta y = \Delta\hat{x}.$$

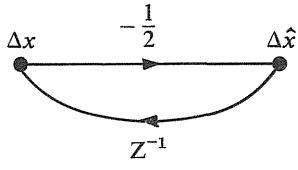
At the true solution, N and y may be expressed in terms of x, giving either of these two results. Since the loop transmittance has a magnitude that is greater than 1, this process





is *unstable* and the error Δx is magnified by -2 with each iteration.

By a similar analysis, you should be able to show that for the second scheme, the linearized graph is as shown, and the error on successive iterations is *decreased* in mag-



nitude by $\frac{1}{2}$ and *reversed* in sign. With successive iterations, the error Δx dies away to zero.

The fact that the successive errors are of opposite sign and decreased by a constant factor suggests that by combining two successive estimates of x in the proper proportion, the errors may be made to cancel each other. Can you design a numerical process that does this?

QUESTION 9.9 The distances in Figure 9.8 have been drawn for $s_0 = 1 + j1$. Show that at s_0 the value of the polynomial represented in Figure 9.8 is

$$\sqrt{1300}/\tan^{-1}(1/3) + \tan^{-1}(-1/2) + \tan^{-1}(3/2) + \tan^{-1}(-1/1).$$

Express this value numerically in rectangular form.

QUESTION 9.10 Show that the polynomial represented by Figure 9.8 may also be written as

$$P(s) = s^4 + 2s^3 + s^2 - 8s - 20. \tag{9.11}$$

Then use the remainder-theorem method for evaluating polynomials discussed in Chapter 7 to show that

$$P(s_0) = -2s_0 - 34 \tag{9.12}$$

for $s_0 = 1 + j1$. Hence,

$$P(1 + j1) = -36 - j2. \tag{9.13}$$

Comment on which of these two methods you found easier to use numerically. Which of these two methods provides the greater insight into the behavior of $P(s)$ as s is varied?

(Answers given above)

QUESTION 9.11 With reference to the polynomial of Figure 9.8:

1. By simply estimating the lengths and angles by eye, sketch as a function of ω the magnitude of $P(s_0)$ as the point s_0 moves along the $j\omega$-axis.
2. Also, sketch the variation of the angle of $P(s_0)$ as a function of ω (for $\sigma = 0$). What value does this angle approach as $\omega \to \infty$?
3. Sketch the magnitude of $P(s_0)$ as the point moves along the σ-axis. (That is, sketch $P(\sigma + j0)$ as a function of σ.)
4. Also sketch the angle of $P(\sigma + j0)$ as a function of σ.
5. Now verify your sketches by calculating the values of $P(s_0)$ at a few selected points. (Answer)

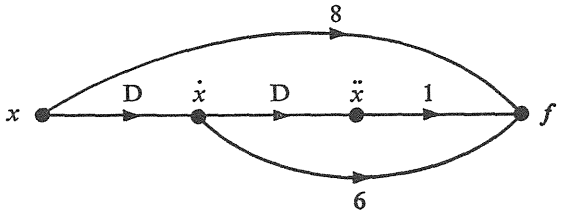
In subsequent studies, you will find this graphical technique for visualizing the behavior of a polynomial extremely useful. It also provides insight into the behavior of more complicated functions, as illustrated by Question 9.10. All the transmittance functions discussed in this section have a well-defined value for any finite value of s. Furthermore, they can be expanded in a Taylor's series about any finite point in the s-plane and are therefore *analytic* through the entire s-plane. Such functions are called *entire functions*. Polynomials comprise the simplest and most important subclass of entire functions; but observe that e^{-st} is also an entire function, and there are others.

The transmittance functions of most physical operators are almost never entire functions. They are well behaved for all values of s in the *right half* of the s-plane (corresponding to the region of physical measurement, as we have seen), but when these functions are *analytically continued* into the left half of the s-plane, one usually finds certain values of s, called *singular points*, at which the value of the function is not well defined. To see how these singular points can arise, carefully consider the next question.

QUESTION 9.12 An external force f and displacement x of a mechanical system are known to be related by the differential equation

$$\ddot{x} + 6\dot{x} + 8x = f. \tag{9.14}$$

Treating the displacement x as an independent or source signal, the external force required to produce that displacement is described by the operator graph as shown.



1. What is the transmittance function that would yield f if the displacement x were a general exponential g_s?
2. Write an expression for the value of the force at any instant t if $x = [10/\underline{0 \text{ deg}}]g_s$ and $s = 1 + j2$. What are the phasors that describe x and f in this case?

3. By examining the transmittance function found in part 1, can you find any values of s for which the force is *null*? What physical interpretation can you give to these zeros of the transmittance function?
4. Suppose that you wish to treat the force as the *independent* signal and the displacement x as *dependent*. *Invert* an appropriate path in the operator graph so as to obtain a new graph, containing only *accumulators* and *scalors*, with f the source node. If f is a general exponential g_s, what will be the displacement x?
5. What is the value of the transmittance function of the new operator obtained in part 4 at any of the zeros found in part 3?

The Poles of a Transmittance Function

An important class of transmittance functions, associated with lumped physical systems that use only accumulators and scalors, may be written as the ratio of two polynomials in s. This is the class of *rational functions*:

$$H(s) = \frac{a_0 s^m + a_1 s^{m-1} + \cdots + a_{m-1} s + a_m}{s^n + b_1 s^{n-1} + \cdots + b_n s + b_n} = \frac{A(s)}{B(s)}, \tag{9.15}$$

where $A(s)$ and $B(s)$ are polynomials in s of degrees m and n, respectively. The rational function is said to be *proper* if the degree m of the numerator is less than the degree n of the denominator. An *improper* rational function may always be expressed as a suitable polynomial in s of degree $m - n$ plus a proper rational functional merely by dividing B into A and writing the quotient and remainder terms (exactly as in Equation 4.10 of Chapter 4). Hence, we focus attention primarily on *proper rational functions.*

It is evident that $H(s)$ will be zero whenever $A(s)$ is zero (provided, of course, that $B(s)$ is not simultaneously zero and the form indeterminant). Thus, the *zeros of a rational function are identical to the zeros of its numerator.* But now a new property, not found in *entire* functions, is encountered. The denominator polynomial $B(s)$ will likewise become zero for certain values of s. At those points, the magnitude of $H(s)$ becomes infinitely large.

Imagine a tent made of very elastic canvas that is erected over the s-plane in such a way that the surface of the stretchy canvas is elevated above the s-plane at every point by distance equal to the *magnitude* of $H(s)$ at that point. Then at the position of the zeros, the tent will appear to be nailed to the ground, whereas in the vicinity of each point where the magnitude becomes infinitely large, the tent will appear to be supported by a pole. Hence, those values of s for which $B(s)$ vanishes and $H(s)$ becomes infinitely large are called the *poles* of the rational function $H(s)$.

Evidently, the monic polynomial $B(s)$ appearing in Equation 9.15 is completely described by its zeros, which, as we have seen, define the *poles* of $H(s)$. Denoting the kth pole of $H(s)$ by p_k, $H(s)$ may then be written in factored form as

$$H(s) = a_0 \frac{(s - z_1)(s - z_2) \cdots (s - z_m)}{(s - p_1)(s - p_2) \cdots (s - p_n)}. \tag{9.16}$$